

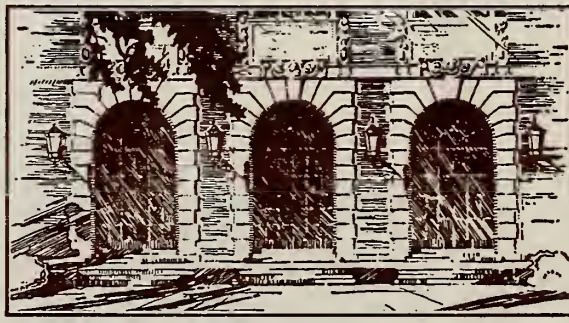
LIBRARY OF THE
UNIVERSITY OF ILLINOIS
AT URBANA-CHAMPAIGN

510.84

Il6r

no.487-492

cop.2



The person charging this material is responsible for its return to the library from which it was withdrawn on or before the **Latest Date** stamped below.

Theft, mutilation, and underlining of books are reasons for disciplinary action and may result in dismissal from the University.

UNIVERSITY OF ILLINOIS LIBRARY AT URBANA-CHAMPAIGN

SEP 21 1973
DEC 27 Recd

510.84
Ill 62
no. 492
cop. 2

UIUCDCS-R-71-492

Math

AN ECONOMICAL ALGORITHM
FOR THE SOLUTION OF FINITE DIFFERENCE EQUATIONS

by

Martin A. Diamond

December 1971



DEPARTMENT OF COMPUTER SCIENCE
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN · URBANA, ILLINOIS

THE LIBRARY OF THE

MAT 2 1972

UNIVERSITY OF ILLINOIS
AT URBANA-CHAMPAIGN

AN ECONOMICAL ALGORITHM FOR THE
SOLUTION OF ELLIPTIC DIFFERENCE EQUATIONS
INDEPENDENT OF USER-SUPPLIED PARAMETERS

BY

MARTIN ALLEN DIAMOND

B.A., Arizona State University, 1967

M.A., Arizona State University, 1968

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Computer Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 1972

Urbana, Illinois



Digitized by the Internet Archive
in 2013

<http://archive.org/details/economicalgori492diam>

AN ECONOMICAL ALGORITHM FOR THE SOLUTION
OF ELLIPTIC DIFFERENCE EQUATIONS INDEPENDENT OF USER-SUPPLIED PARAMETERS *

by

Martin Diamond

December 5, 1971
Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, Illinois 61801

* This work was supported in part by the National Science Foundation under Grant No. GJ812 and was submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Computer Science, February, 1972.

ACKNOWLEDGEMENT

The author wishes to express his appreciation and gratitude to Professor P. E. Saylor for his help in editing which has greatly improved the readability of this thesis.

The author would also like to thank Professor D. B. Gillies for his financial support. This work was supported in part by the National Science Foundation under Grant No. GJ812.

Finally, the author wishes to thank Mrs. Connie Slovak for her aid in preparing the manuscript.

TABLE OF CONTENTS

	Page
1. INTRODUCTION.	1
2. PRELIMINARIES	6
2.1 Origin Of The Problem	6
2.2 Known Results.	10
3. CONVERGENCE CRITERIA.	13
3.1 Convergence Criterion When A And A+B Are Arbitrary	13
3.2 The Positive Definite Case	15
3.3 The Use Of A Sequence of τ 's When A+B Is Positive Definite . .	17
3.4 Obtaining A Sequence of τ 's When A+B Is Not Positive Definite.	20
4. USING A SEQUENCE OF τ 's WHEN LITTLE IS KNOWN ABOUT THE EIGENVALUES OF $(A+B)^{-1}A$	31
4.1 Introduction	31
4.2 Calculating An Eigenvector and Eigenvalue of $(A+B)^{-1}A$	34
4.3 Application Of The Method.	41
4.4 Using The Recursive Property Of The Tchebychev Polynomials . .	43
4.5 The Algorithm And Its Convergence.	53
4.6 Experimental Results	61
5. AN ESTIMATE OF $\lambda_{\min}$	78
5.1 The Limiting Values Of $\hat{b}_{i,j}$	78
5.2 An Estimate Of $\lambda_{\min}$	91
LIST OF REFERENCES.	96
VITA.	98

1. INTRODUCTION

The principal result of this thesis is the development of an algorithm to solve the large sets of linear equations that arise in the solution of elliptic partial differential equations by implicit numerical techniques. The algorithm is an important development for the following reasons:

1. The algorithm is always convergent.
2. The algorithm requires no set of parameters.
3. The algorithm is not dependent upon special properties of the partial differential equation or the matrix of coefficients of the resulting linear equations.
4. The algorithm has proved to be rapidly convergent on a variety of test problems.*
5. The use of this algorithm in conjunction with a factorization due to Y. J. Kim [11] promises to be the first effective iterative method to solve the systems of equations generated in the application of finite element techniques.

The algorithm developed in this thesis applied in conjunction with a factorization is the only existing technique with all of these properties.

*Although a priori estimates of the rate of convergence can be established, they depend upon the extreme eigenvalues of the iteration matrix. If accurate approximations of these eigenvalues are available and are used in the algorithm, then the algorithm reduces to the method of Dupont, Kendall and Rachford [5] with optimal parameters. Thus in this case the rate of convergence can be computed as described in [5]. The application of the algorithm, however, assumes no knowledge of the eigenvalues of the iteration matrix. Therefore no a priori bound for the rate of convergence is computable.

Since the application of direct methods to the large systems we are considering is either impossible or impractical, iterative methods are usually used. Unfortunately the standard iterative techniques are also deficient. The Jacobi method is quite slow even on some simple problems. The Gauss-Seidel [19] method can be shown to be simply a factor of two better than the Jacobi iteration regardless of the number of equations. Successive-Over-Relaxation [19] is dependent on the problem and may become quite slow.

The Alternating Direction Implicit iteration [20] is more efficient than either Gauss-Seidel or Successive-Over-Relaxation. However, it requires the selection of a set of iteration parameters to be applied cyclically during the iteration. In general, there is no theoretical basis for selecting a set that will give rapid convergence. Moreover, the method loses its fast convergence when the region of definition of the boundary value problem is irregular.

The algorithm developed in this thesis is based on factorization techniques. To elucidate the difference between this algorithm and other algorithms based on factorizations, we shall give a brief explanation of factorization methods.

The solution of a system of equations $Ax = q$ by factorization techniques can be separated into two segments. The first step is the definition of an auxiliary matrix, B , which is chosen so that $A+B = LU$ where L and U are sparse lower and upper triangular matrices respectively. In practical computer programs the elements of B are not computed; instead the elements of L and U are computed. Then by using the factors L and U , systems of equations whose matrix of coefficients is $A+B$ can be easily solved.

The second part of the method consists of an algorithm which uses an iteration of the form

$$(A+B)x_{n+1} = (A+B)x_n - (Ax_n - q), \quad n = 0, 1, 2, \dots$$

to generate a sequence of vectors $\{x_n\}$ which converges to the solution of $Ax = q$. The convergence of the iteration is dependent upon the definition of the auxiliary matrix (through the definition of L and U) and a sequence of parameters.

One of two types of parameters is employed in the iteration. The first type may be used to alter the auxiliary matrix for each n . There has been no analysis of the resulting iteration,

$$(A+B_n)x_{n+1} = (A+B_n)x_n - (Ax_n - q).$$

Therefore there is no theoretical basis for the selection of such a set of parameters.

The second set of parameters, denoted by τ_n , multiplies the term $(Ax_n - q)$ in the iteration as follows

$$(A+B)x_{n+1} = (A+B)x_n - \tau_n(Ax_n - q).$$

In previous algorithms the definition of the parameters τ_n is based upon a lower bound for $\min_{x \neq 0} \frac{\langle Ax, x \rangle}{\langle (A+B)x, x \rangle}$ and an upper bound for $\max_{x \neq 0} \frac{\langle Ax, x \rangle}{\langle (A+B)x, x \rangle}$; equivalently they are based on lower and upper bounds for the extreme eigenvalues of $(A+B)^{-1}A$. These bounds must be calculated every time the matrix of coefficients A , is altered. Furthermore, the rates of convergence of the previous algorithms are dependent upon the accuracy of the bounds. The new algorithm requires no knowledge of the eigenvalues of $(A+B)^{-1}A$. Moreover the rate of convergence is improved by quantities calculated during the execution of the algorithm in such a way that the rate of convergence approaches the optimal rate.

The new algorithm is also superior to algorithms which vary the auxiliary matrix for each iteration. Its primary advantage is that there is a mathematical basis which guarantees convergence. A second advantage is that only one auxiliary matrix is used. Thus only one pair of factors L and U is defined and consequently less work is required.

Although no knowledge of the eigenvalues of $(A+B)^{-1}A$ is necessary, information about the eigenvalues is useful. Furthermore, the rate of convergence is governed by the extreme eigenvalues. Therefore the selection of the auxiliary matrix is still important. For matrices which arise in the solution of elliptic partial differential equations by finite difference methods, the symmetric auxiliary matrix defined by Stone [15] (see section 2.2) seems to be the best choice of the known factorizations.

We shall now summarize the results of this thesis.

In Chapter 3 it is shown that for arbitrary A and $A+B$ the iteration

$$(A+B)x_{n+1} = (A+B)x_n - \tau_n (Ax_n - q)$$

is convergent with τ_n a constant τ iff (if and only if) the real part of each of the eigenvalues of $A^{-1}B$ is greater than $\frac{\tau}{2} - 1$. By modifying this result for positive definite A and Hermitian B , it is shown that the iteration is convergent with $\tau_n \equiv 1$ iff $A+2B$ is positive definite. This strengthens the long established result that when A is positive definite and B is Hermitian, the iteration is convergent with $\tau_n \equiv 1$ if $A+2B$ is positive definite, see Wachspress [20]. Next it is shown that if A and $A+B$ are positive definite then the iteration is convergent with a constant value τ iff

$$0 < \tau < \frac{2}{\max_i \lambda_i((A+B)^{-1}A)} ,$$

where $\lambda_i((A+B)^{-1}A)$ is an eigenvalue of $(A+B)^{-1}A$.

In section 3.3 the analysis by Dupont, Kendall and Rachford [5] concerning the selection of a best sequence of τ_n 's to use when the eigenvalues of $(A+B)^{-1}A$ are real is presented. The results are extended in section 3.4 establishing a best sequence of τ_n 's to use when the eigenvalues of $(A+B)^{-1}A$ are complex.

In Chapter 4 an algorithm based on the iteration

$$(A+B)x_{n+1} = (A+B)x_n - \tau_n (Ax_n - q)$$

is developed.

In Chapter 5 an approximation of the smallest eigenvalue of $(A+B_S)^{-1}A$, where B_S is the symmetric auxiliary matrix defined by Stone, is derived for some special cases. This approximation can be used in starting the algorithm of Chapter 4 when B_S is the auxiliary matrix.

2. PRELIMINARIES.

Notation: Throughout this thesis, $\lambda_i(M)$ will denote an eigenvalue of the matrix M , $\lambda_{\min}(M)$ will denote the eigenvalue of minimum algebraic value, $\lambda_{\max}(M)$ will denote the eigenvalue of largest algebraic value, and $\rho(M)$ will be used to denote the spectral radius of M , i.e., $\rho(M) = \max_i |\lambda_i(M)|$.

All matrices used in this thesis, even those specified as general, are real matrices.

2.1 Origin Of The Problem. Consider the Dirichlet problem

$$-\frac{\partial}{\partial x_1} (a_1(x) \frac{\partial u}{\partial x_1}) - \frac{\partial}{\partial x_2} (a_2(x) \frac{\partial u}{\partial x_2}) = q(x), \quad x \in D, \quad u(x) = (x), \quad \text{for } x \in \partial D \quad (2.1)$$

where $x = (x_1, x_2)$, D is the interior of a compact region with boundary ∂D and where the differential operator in (2.1) is elliptic, that is $a_1(x)$ and $a_2(x)$ are strictly positive on D . A common method of solving (2.1) is to approximate the differential operator with a difference operator. A system of linear equations results and the problem is then to find a solution to the system of equations.

The system is derived as follows. Assume D lies in the first quadrant; cover D with two families of parallel lines,

$$x_1 = jh \quad j = 1, \dots, n_1$$

and

$$x_2 = kh \quad k = 1, \dots, n_2.$$

Denote the points of intersection (jh, kh) by (j, k) . The points (j, k) are called the mesh points or gridpoints. Let ∂D_h be the polygon formed by

joining those mesh points (j,k) for which one of $(j\pm 1,k)$ or $(j,k\pm 1)$ is not in D . Similarly let D_h be the open and connected set bounded by ∂D_h . At each gridpoint of D_h the differential operator is approximated by a difference operator and the equations are in one-to-one correspondence with the gridpoints.

For simplicity of notation we shall consider the region D to be the unit square bounded by the x_1 and x_2 axes and the lines $x_1=1$ and $x_2=1$. Let $x = (x_{1,1}, x_{1,2}, \dots, x_{n,1}, x_{2,1}, \dots, x_{n,n})^T$ where $n_1 = n_2 = n$. We obtain the matrix equation

$$Ax = q \quad (2.2)$$

defined by $(Ax)_{j,k} = q_{j,k}$, where the subscript corresponds to gridpoint (j,k) and $q_{j,k}$ is derived from $q(jh,kh)$ and the boundary conditions.

The difference operator can be chosen so that matrix A is diagonally dominant and positive definite with $a_{i,i} > 0$ and $a_{i,j} \leq 0$ for $i \neq j$. One method for deriving such an A is to approximate $\frac{\partial}{\partial x_i} (a_i(x) \frac{\partial}{\partial x_i} u(x))$ by

$$h^{-2} \{ a_i(x + \frac{h}{2} e_i) [u(x) - u(x + h e_i)] + a_i(x - \frac{h}{2} e_i) [u(x) - u(x - h e_i)] \}$$

where e_1 and e_2 are unit vectors along the x_1 and x_2 axes respectively. This defines A by

$$(Ax)_{j,k} = B_{j,k} x_{j,k-1} + D_{j,k} x_{j-1,k} + E_{j,k} + D_{j+1,k} x_{j+1,k} + B_{j,k+1} x_{j,k+1},$$

$j,k = 1, \dots, n$, where

$$B_{j,k} = -a_2(jh, (k - \frac{1}{2})h),$$

$$D_{j,k} = -a_1((j - \frac{1}{2})h, kh), \quad (2.3)$$

and

$$E_{j,k} = a_2(jh, (k - \frac{1}{2})h) + a_2(jh, (k + \frac{1}{2})h) + a_1((j - \frac{1}{2})h, kh) \\ + a_1((j + \frac{1}{2})h, kh).$$

When $j=1$ or $k=1$ we define $D_{j,k} = 0$ or $B_{j,k} = 0$ respectively.

The following properties of A and the quantities defined in (2.3) will be used below.

$$(Ax)_{j,k} = E_{j,k}x_{j,k} + B_{j,k}x_{j,k-1} + D_{j,k}x_{j-1,k} + D_{j+1,k}x_{j+1,k} + B_{j,k+1}x_{j,k+1},$$

$$B_{j,k} \leq 0, \quad B_{j,1} = 0,$$

$$D_{j,k} \leq 0, \quad D_{1,k} = 0, \tag{2.4}$$

and

$$E_{j,k} \geq -2(B_{j,k} + D_{j,k}) \text{ with equality holding if } B_{j,k} \neq 0 \text{ and } D_{j,k} \neq 0$$

for $j=1, 2, \dots, n$ and $k=1, 2, \dots, n$.

The subscript (j,k) refers to the gridpoint (jh, kh) .

In this thesis we shall study the solution of equations

$$Ax = q$$

by factorization techniques. As described in the introduction, the idea of a factorization method is to add an auxiliary matrix, B , to A so that $A+B$ factors into known sparse matrices L and U . An algorithm based on the iteration

$$(A+B)x_{n+1} = (A+B)x_n - \tau_n(Ax_n - q) \tag{2.5}$$

is then used to compute a sequence of vectors $\{x_n\}$ which converges to the solution x . The parameters τ_n are defined by the algorithm.

A factorization is defined by the auxiliary matrix or the factors L and U. We shall give particular attention to the symmetric factorization due to Stone [15]. This factorization which can be applied to any matrix A with the properties listed in (2.4), is defined as follows.

$$\begin{aligned} (Lx)_{j,k} &= b_{j,k}x_{j,k-1} + c_{j,k}x_{j-1,k} + d_{j,k}x_{j,k} , \\ (Ux)_{j,k} &= x_{j,k} + e_{j,k}x_{j+1,k} + f_{j,k}x_{j,k+1} , \end{aligned} \quad (2.6)$$

where

$$\begin{aligned} b_{j,k} &= B_{j,k} - \alpha c_{j,k-1} f_{j-1,k-1} , \\ c_{j,k} &= D_{j,k} - \alpha b_{j-1,k} e_{j-1,k-1} , \\ d_{j,k} + b_{j,k} f_{j,k-1} + c_{j,k} e_{j-1,k} &= E_{j,k} + \alpha c_{j,k-1} f_{j-1,k-1} \\ &\quad + \alpha b_{j-1,k} e_{j-1,k-1} , \\ d_{j,k} e_{j,k} &= D_{j+1,k} - \alpha b_{j,k} e_{j,k-1} , \end{aligned} \quad (2.7a)$$

and

$$d_{j,k} f_{j,k} = B_{j,k+1} - \alpha c_{j,k} f_{j-1,k} .$$

The quantities $B_{j,k}$, $D_{j,k}$ and $E_{j,k}$ are the elements of A defined in (2.4), and the parameter α is used to vary the auxiliary matrix if Stone's algorithm [16] based on the iteration (2.5) is used. We are interested in algorithms that don't vary the auxiliary matrix and will assume $\alpha = 1$.

In addition to the quantities defined in (2.7a) we define

$$C_{j,k} = b_{j,k} e_{j,k-1} \quad \text{and} \quad G_{j,k} = c_{j,k} f_{j-1,k} . \quad (2.7b)$$

Since there should be no confusion as to whether the auxiliary matrix is defined by (2.6) and (2.7) or is arbitrary, B will be used to denote either.

With L and U of the form (2.6), A+B is given by

$$\begin{aligned}
 ((A+B)x)_{j,k} &= ((LU)x)_{j,k} = b_{j,k}x_{j,k-1} + b_{j,k}e_{j,k-1}x_{j+1,k-1} \\
 &+ c_{j,k}x_{j-1,k} + (d_{j,k} + b_{j,k}f_{j,k-1} + c_{j,k}e_{j-1,k})x_{j,k} \\
 &+ d_{j,k}e_{j,k}x_{j+1,k} + c_{j,k}B_{j-1,k}x_{j-1,k+1} + d_{j,k}f_{j,k}x_{j,k+1}.
 \end{aligned} \tag{2.8}$$

2.2 Known Results The following definitions are standard.

Definition 1.1: If $x=(x_1, \dots, x_n)^T$ and $y=(y_1, \dots, y_n)^T$ are n-component

vectors then $\langle x, y \rangle = \sum_{i=1}^n x_i y_i^*$ where y_i^* denotes the complex conjugate of

y_i . $\langle x, y \rangle$ is called the innerproduct of x and y.

Definition 1.2: $\|x\|$, the L_2 norm of x, is defined by $\|x\| = \|x\|_{L_2} = \langle x, x \rangle^{\frac{1}{2}}$.

If A is positive definite then the A-norm of x, $\|x\|_A$, is defined as $\langle Ax, x \rangle^{\frac{1}{2}}$.

Dupont, Kendall and Rachford [4] have proved the following lemma.

Lemma 2.1: Let A and A+B be positive definite. Suppose there exist positive numbers e_1 and e_2 such that for all nonzero x we have

$$\frac{\langle Ax, x \rangle}{\langle (A+B)x, x \rangle} \in [e_1, e_2] \tag{2.9}$$

Then for $0 < \tau < 2e_2^{-1}$ the sequence $\{x_n\}$ defined by (2.5) converges to the solution of $Ax=q$. Furthermore, for $\tau \equiv 2(e_1 + e_2)^{-1}$ the A-norm of the error,

$\epsilon_n = x - x_n$, is reduced by a factor of at least $(e_1 - e_2)/(e_1 + e_2)$ in each iteration.

It is also suggested [4] that a sequence of τ_n derived from a classical Tchebychev analysis be used. Iteration (2.5) is then performed in practice by using a technique due to Stiefel [4]. This technique will be examined in Chapter 3 and a modification of it is used in Chapter 4 in developing a new algorithm.

Bracha [1] has derived the following important properties of the elements of the factors L and U defined in (2.6) and (2.7).

$$(a) \quad b_{j,k} \leq B_{j,k} \leq 0,$$

$$(b) \quad c_{j,k} \leq D_{j,k} \leq 0,$$

$$(c) \quad d_{j,k} > 0,$$

$$(d) \quad -1 < e_{j,k} \leq 0,$$

$$(e) \quad -1 < f_{j,k} \leq 0,$$

and

$$(f) \quad 1 + e_{j,k} + f_{j,k} > 0.$$

Bracha has also shown in [1] that for the factorization defined in (2.6) and (2.7) $A+B$ is positive definite when A satisfies (2.4). A more concise proof is included here.

Lemma 2.2: If A satisfies (2.4) and B is defined by (2.6) and (2.7) then $A+B$ is positive definite.

Proof: From (2.7) it is seen that $b_{j,k} = d_{j,k-1} f_{j,k-1}$ and $c_{j,k} = d_{j-1,k} e_{j-1,k}$.

Thus (2.6) implies $(L)_{r,s} = d_{j,k} \cdot (U)_{s,r}$ where row r of $A+B$ corresponds to

the gridpoint (j,k) and $s=1, \dots, r$. We have $L = U^T \text{diag}(d_{j,k})$, where $\text{diag}(d_{j,k})$ is the diagonal matrix with $d_{j,k}$ along the main diagonal and U^T is the transpose of U . Using property (c) from above, we have

$$A+B = LU = U^T \text{diag}(d_{j,k}^{1/2}) \text{diag}(d_{j,k}^{1/2}) U$$

which is clearly positive definite.

In addition Bracha proved the following theorem [2].

Theorem 2.1: Let A and $A+B$ be the matrices defined in (1.6) and (1.10) respectively. Let $e_{j,k}$ and $f_{j,k}$ be defined by (1.10). Let

$$\beta_{j,k} = \frac{1}{-e_{j,k} - f_{j,k}} \quad j,k = 1, 2, \dots, n,$$

and

$$\beta = \min_{j,k} \beta_{j,k}.$$

Then

$$\max_{x \neq 0} \frac{\langle Ax, x \rangle}{\langle (A+B)x, x \rangle} \leq \frac{1}{1 - \frac{1}{\beta}}$$

and

$$\min_{x \neq 0} \frac{\langle Ax, x \rangle}{\langle (A+B)x, x \rangle} \geq \frac{1}{1 + k_1}$$

where k_1 is a positive constant.

3. CONVERGENCE CRITERIA

In this chapter necessary and sufficient conditions are derived for the convergence of $(A+B)x_{n+1} = (A+B)x_n - \tau_n(Ax_n - q)$. The conditions we obtain do not seem to be known in the literature. The conditions are first developed for arbitrary A and B . For this case it is shown that the iteration is convergent with a constant τ iff τ is selected from a certain region in the complex plane. This result is modified for the case when A is positive definite and further modified for the case when both A and $A+B$ are positive definite. When A and $A+B$ are positive definite, a best constant value of τ is established.

Section 3.4 establishes a best sequence of τ_n 's to use when the eigenvalues of $(A+B)^{-1}A$ are complex. The derivation follows that of the known results presented in Section 3.3, which establish a best sequence of τ_n 's when the eigenvalues of $(A+B)^{-1}A$ are real.

3.1 Convergence Criterion When A and $A+B$ Are Arbitrary

In the following two lemmas convergence criteria are derived in terms of τ for $(A+B)x_{n+1} = (A+B)x_n - \tau(Ax_n - q)$. (3.1)

Lemma 3.1: Iteration (3.1) is convergent iff $\operatorname{Re}(\lambda_1(A^{-1}B)) > \frac{\tau}{2} - 1$.

Proof: Let x be the exact solution of $Ax=q$. Let the error in the n^{th} iterate be given by $e_n = x - x_n$. Subtracting (3.1) from

$$(A+B)x = (A+B)x - \tau(Ax - q)$$

yields $(A+B)(x - x_{n+1}) = (A+B)(x - x_n) - \tau(A(x - x_n))$

or
$$e_{n+1} = (A+B)^{-1}(A+B - \tau A)e_n.$$

The matrix $(A+B)^{-1}(A+B-\tau A)$ is the iteration matrix of the iteration (3.1).

It is well known that the iteration is convergent iff

$$\rho((A+B)^{-1}(A+B-\tau A)) < 1. \quad (3.2)$$

The iteration matrix can be rewritten as

$$\begin{aligned} (A+B)^{-1}(A+B-\tau A) &= (I-\tau(A+B)^{-1}A) \\ &= (I-\tau(A^{-1}(A+B))^{-1}) = (I-\tau(I+A^{-1}B)^{-1}), \end{aligned}$$

where I is the identity matrix.

Thus (3.2) can be rewritten as

$$\lambda_i(I-\tau(I+A^{-1}B)^{-1}) \in \delta(0,0;1)$$

where

$$\delta(u,v;R) = \{(u,v):(u-x)^2+(v-y)^2 < R^2\}$$

Hence (3.1) is convergent iff

$$\begin{aligned} \lambda_i\left(\frac{1}{\tau}I-(I+A^{-1}B)^{-1}\right) &\in \delta\left(0,0;\frac{1}{\tau}\right), \\ \lambda_i\left((I+A^{-1}B)^{-1}\right) &\in \delta\left(\frac{1}{\tau},0;\frac{1}{\tau}\right), \end{aligned} \quad (3.3)$$

or $\lambda_i(I+A^{-1}B)$ has real part greater than $\tau/2$. It now follows immediately that (3.1) is convergent iff

$$\operatorname{Re}(\lambda(A^{-1}B)) > \frac{\tau}{2} - 1. \quad (3.4)$$

Corollary 3.1.1: If A is positive definite and B is Hermitian, then (3.1) is convergent iff $A+B-\frac{\tau}{2}A$ is positive definite.

Proof: If A and B are Hermitian, each $\lambda_i(A^{-1}B)$ is real and (3.4) becomes

$$\lambda_i(A^{-1}B) > \frac{\tau}{2} - 1.$$

This is equivalent to $\lambda_i(I(1-\frac{\tau}{2})+A^{-1}B) > 0$. But

$$I(1-\frac{\tau}{2})+A^{-1}B = A^{-1}(A(1-\frac{\tau}{2})+B) = A^{-1}(A+B-\frac{\tau}{2}A).$$

Thus (3.4) is equivalent to $\lambda_1(A^{-1}(A+B - \frac{\tau}{2}A)) > 0$. (3.5)

Since $A^{-1}(A+B - \frac{\tau}{2}A)$ is similar to $A^{-1/2}(A+B - \frac{\tau}{2}A)A^{-1/2}$,

(3.5) may be restated as

$$\langle A^{-1/2}(A+B - \frac{\tau}{2}A)A^{-1/2}x, x \rangle > 0, \text{ for all nonzero vectors } x.$$

Therefore

$$\langle (A+B - \frac{\tau}{2}A)A^{-1/2}x, A^{-1/2}x \rangle > 0, \text{ for all nonzero } x.$$

Letting $y = A^{-1/2}x$ we obtain $\langle (A+B - \frac{\tau}{2}A)y, y \rangle > 0$, for all nonzero y .

Hence (3.1) is convergent iff $A+B - \frac{\tau}{2}A$ is positive definite, and this completes the proof.

For $\tau = 1$ the corollary implies (3.1) converges iff $A+2B$ is positive definite. This strengthens the previously known fact that a sufficient condition for the convergence of (3.1) is that $A+2B$ be positive definite.

3.2 The Positive Definite Case

We now modify the above results for an $A+B$ which is positive definite, as it is in the factorization due to Dupont, Kendall and Rachford [5] and the symmetric factorization due to Stone [15].

Lemma 3.2: If A and $A+B$ are positive definite then (3.1) is convergent iff $0 < \tau < \frac{2}{\lambda_{\text{MAX}}((A+B)^{-1}A)}$.

Proof: (3.1) is convergent iff

$$|\lambda_1(I - \tau(A+B)^{-1}A)| < 1.$$

Since $\lambda_1((A+B)^{-1}A) > 0$, this condition is equivalent to

$$1 - \tau \lambda_{\text{MAX}}((A+B)^{-1}A) > -1$$

or $0 < \tau < \frac{2}{\lambda_{\text{MAX}}((A+B)^{-1}A)}$, and this completes the proof.

We can make a further refinement by deriving a best value for τ as seen in the next lemma.

Lemma 3.3: If A and $A+B$ are positive definite then

$$\tau_b = \frac{2}{\lambda_{\text{MAX}}((A+B)^{-1}A) + \lambda_{\text{min}}((A+B)^{-1}A)}$$

is the best choice for τ in (3.1).

Proof: The iteration matrix of (3.1) is $T_\tau = (I - \tau(A+B)^{-1}A)$. Thus, since all the eigenvalues of $(A+B)^{-1}A$ are positive, the eigenvalues of T_τ are decreasing as a function of τ . Clearly the spectral radius of T_τ is minimized with respect to τ , when $|\lambda_{\text{min}}(T_\tau)| = |\lambda_{\text{MAX}}(T_\tau)|$, i.e.,

$$|1 - \tau \lambda_{\text{MAX}}((A+B)^{-1}A)| = |1 - \tau \lambda_{\text{min}}((A+B)^{-1}A)|,$$

or
$$2 = \tau [\lambda_{\text{MAX}}((A+B)^{-1}A) + \lambda_{\text{min}}((A+B)^{-1}A)].$$

Therefore $\tau_b = 2/[\lambda_{\text{MAX}}((A+B)^{-1}A) + \lambda_{\text{min}}((A+B)^{-1}A)]$ is the best choice of τ as claimed.

Moreover for $\tau = \tau_b$, $\rho(T_\tau) = 1 - \frac{2\lambda_{\text{min}}((A+B)^{-1}A)}{\lambda_{\text{MAX}}((A+B)^{-1}A) + \lambda_{\text{min}}((A+B)^{-1}A)}.$

There is a strong relation between Lemmas 3.2 and 3.3 and Lemma 2.1. Lemma 3.2 states a necessary and sufficient condition while Lemma 2.1 states only a sufficient condition. The hypothesis of Lemma 2.1, however, includes a supposition that positive numbers e_1 and e_2 exist such that for all nonzero x

$$\frac{\langle Ax, x \rangle}{\langle (A+B)x, x \rangle} \in [e_1, e_2].$$

This is certainly valid when both A and $A+B$ are positive definite.

In fact

$$e_1 = \frac{\lambda_{\text{min}}(A)}{\lambda_{\text{MAX}}(A+B)}$$

and

$$e_2 = \frac{\lambda_{\text{MAX}}(A)}{\lambda_{\text{min}}(A+B)}$$

are sufficient. The next lemma shows precisely the connection of Lemmas 2.1 and 3.2.

Lemma 3.4: Let A and $A+B$ be positive definite. Then

$$\frac{\langle Ax, x \rangle}{\langle (A+B)x, x \rangle} \in [e_1, e_2] \quad \text{for all nonzero } x$$

iff $\lambda_i((A+B)^{-1}A) \in [e_1, e_2]$.

Proof: The conclusion may be restated as

$$\lambda_{\text{min}}((A+B)^{-1}A) = \min_{y \neq 0} \frac{\langle ay, y \rangle}{\langle (A+B)y, y \rangle} \quad (3.6)$$

and

$$\lambda_{\text{MAX}}((A+B)^{-1}A) = \max_{y \neq 0} \frac{\langle ay, y \rangle}{\langle (A+B)y, y \rangle}.$$

Now since $(A+B)^{-1}A$ is similar to $A^{1/2}(A+B)^{-1}A^{1/2}$ which is Hermitian (3.7)

$$\lambda_{\text{min}}((A+B)^{-1}A) = \lambda_{\text{min}}(A^{1/2}(A+B)^{-1}A^{1/2}) = \min_{x \neq 0} \frac{\langle A^{1/2}(A+B)^{-1}A^{1/2}x, x \rangle}{\|x\|^2}$$

Letting $x = [A^{1/2}(A+B)^{-1}A^{1/2}]^{-1/2} A^{1/2} y$ and using the selfadjointness

of $[A^{1/2}(A+B)^{-1}A^{1/2}]^{-1/2}$ and $A^{1/2}$ (3.7) becomes $\lambda_{\text{min}}((A+B)^{-1}A) =$

$$\min_{y \neq 0} \frac{\langle Ay, y \rangle}{\langle (A+B)y, y \rangle}.$$

$$\text{Similarly } \lambda_{\text{MAX}}((A+B)^{-1}A) = \max_{y \neq 0} \frac{\langle Ay, y \rangle}{\langle (A+B)y, y \rangle} \quad \text{This completes}$$

the proof.

3.3 The Use Of A Sequence Of τ 's When $A+B$ Is Positive Definite

There are no new results presented in this section. However, Section 3.4 and Chapter 4 are developed from the results described here.

In this section a "best" sequence of τ_n 's to be used in the iteration $(A+B)x_{n+1} = (A+B)x_n - \tau_n(Ax_n - q)$, is derived for the case when both A and $A+B$ are positive definite. A "best" sequence is one that minimizes the spectral radius of the iteration matrix.

As in Lemma 3.1 the error vector e_n is multiplied at each step by $(I - \tau_n(A+B)^{-1}A)$. Hence

$$e_N = \prod_{n=0}^{N-1} (I - \tau_n(A+B)^{-1}A) e_0. \quad (3.8)$$

To make the iteration converge as rapidly as possible, the τ_n 's should be chosen to minimize the spectral radius of

$$M_N = \prod_{n=0}^{N-1} (I - \tau_n(A+B)^{-1}A).$$

The eigenvalues of M_N are

$$\lambda_i(M_N) = \prod_{n=0}^{N-1} [I - \tau_n \lambda_i((A+B)^{-1}A)]$$

and since $\lambda_i((A+B)^{-1}A)$ is real, $\lambda_i(M_N)$ is real. The problem of minimizing the $\lambda_i(M_N)$ thus reduces to selecting a sequence of τ_n 's that minimizes the maximum absolute value of

$$P_N(X) = \prod_{n=0}^{N-1} (1 - \tau_n X) \quad (3.9)$$

on the interval $\lambda_{\min}((A+B)^{-1}A) \leq X \leq \lambda_{\max}((A+B)^{-1}A)$. No normalization is necessary since it is clear that all P_N of the form (3.9) satisfy $P_N(0) = 1$.

By classical Tchebychev analysis it is seen that the desired polynomial is given by

$$P_N(X) = \frac{T_N((\lambda_{\min} + \lambda_{\max} - 2X)/(\lambda_{\max} - \lambda_{\min}))}{T_N((\lambda_{\min} + \lambda_{\max})/(\lambda_{\max} - \lambda_{\min}))}, \quad (3.10)$$

where T_N is the Tchebychev polynomial of degree N and $\lambda_{\min}$ and $\lambda_{\max}$ denote $\lambda_{\min}((A+B)^{-1}A)$ and $\lambda_{\max}((A+B)^{-1}A)$ respectively. The maximum absolute value of $P_N(X)$ on $[\lambda_{\min}, \lambda_{\max}]$ is

$$\left[T_N \left(\frac{\lambda_{\min} + \lambda_{\max}}{\lambda_{\max} - \lambda_{\min}} \right) \right]^{-1}. \quad (3.11)$$

Note that if $\lambda_{\min} \leq 0 < \lambda_{\max}$ then the maximum of P_N must be greater than or equal to 1, since $P_N(0)=1$. However A and $A+B$ positive definite implies that $\lambda_{\min}$ and $\lambda_{\max}$ are positive and therefore assures the reduction of the error vector when $\{\tau_n\}$ is defined by (3.9) and (3.10)

There are two ways of implementing a sequence of τ_n 's so that (3.9) is the polynomial given in (3.10). In the first method an N is selected and the sequence τ_n $n=0,1,\dots,N-1$ is computed as the reciprocal of the roots of $T_N \left(\frac{\lambda_{\min} + \lambda_{\max}}{\lambda_{\max} - \lambda_{\min}} \right)^{-2X}$. The sequence is then used repeatedly until the norm of the residual vector, $r_n = Ax_n - q$, is sufficiently small. A sequence of τ 's derived as above is called a Tchebychev sequence of length N .

The other method of implementing the sequence, which was first suggested by Stiefel [4], uses the recursive property of the Tchebychev polynomials. If this technique is used then every x_n is the result of a Tchebychev sequence of length n . If the first method is used then x_n is the result of a Tchebychev sequence of length n only when $n=N$.

The Stiefel sequence is defined as follows. Suppose x_k is the result of using a Tchebychev sequence of length k . Define x_{k+1} as

$$x_{k+1} = x_k + \delta x_k,$$

where

$$\delta x_k = \frac{4T_k(y)}{(\lambda_{\max} - \lambda_{\min})T_{k+1}(y)} (A+B)^{-1}(q - Ax_k) + \frac{T_{k-1}(y)}{T_{k+1}(y)} \delta x_{k-1},$$

with

$$\delta x_0 = \frac{2}{\lambda_{\min} + \lambda_{\max}} (A+B)^{-1} (q - Ax_0)$$

for

$$Y = \frac{\lambda_{\text{MAX}} + \lambda_{\text{min}}}{\lambda_{\text{MAX}} - \lambda_{\text{min}}} ,$$

then x_{k+1} is the result of using a Tchebychev sequence of length $k+1$.

In general λ_{min} and λ_{MAX} are not known. Tchebychev minimization is typically implemented by finding an interval (e_1, e_2) containing $(\lambda_{\text{min}}, \lambda_{\text{MAX}})$ and using e_1 and e_2 instead of λ_{min} and λ_{MAX} respectively.

In place of (3.9) we have,

$$P_N(X) = \prod_{n=0}^{N-1} (1 - \tau_n X) = \frac{T_N \left[\frac{e_1 + e_2 - 2X}{e_2 - e_1} \right]}{T_N \left[\frac{e_1 + e_2}{e_2 - e_1} \right]} .$$

And the Stiefel sequence is defined by

$$x_{k+1} = x_k + \delta x_k ,$$

$$\delta x_k = \frac{4T_k(y)}{(e_2 - e_1)T_{k+1}(y)} (A+B)^{-1}(q - Ax_k) + \frac{T_{k-1}(y)}{T_{k+1}(y)} x_{k-1} ,$$

$$\delta x_0 = \frac{2}{e_1 + e_2} (A+B)^{-1}(q - Ax_0)$$

where $y = (e_1 + e_2)/(e_2 - e_1)$.

3.4 Obtaining A Sequence Of τ 's When $A+B$ Is Not Positive Definite

In this section we shall derive a best sequence of τ 's to be used in

$$(A+B)x_{n+1} = (A+B)x_n - \tau_n (Ax_n - q)$$

when $(A+B)^{-1}A$ has complex eigenvalues. Before proceeding, it is necessary to define what "best" means. The iteration

$$(A+B)x_{n+1} = (A+B)x_n - \tau_n (Ax_n - q)$$

is most rapidly convergent when $\rho\left(\prod_{n=0}^{N-1} (I - \tau_n (A+B)^{-1}A)\right)$ is minimized.

This is accomplished when

$$\max_i \prod_{n=0}^{N-1} (1 - \tau_n \lambda_i (A+B)^{-1}A) \quad (3.13)$$

is minimized. Since not all of the λ_i are known, the τ_n could be selected to minimize

$$\max_{z \in R} \left| P_N(z) \right|, \quad \text{where } P_N(z) = \prod_{n=0}^{N-1} (1 - \tau_n z) \quad (3.14)$$

and R is some region containing all of the eigenvalues of $(A+B)^{-1}A$.

This concept is formalized with the following definition.

Definition 3.1: A sequence of parameters τ_n $n = 0, 1, \dots, N-1$ used in iteration (2.5) is called a best sequence of length N in a region R containing $\{\lambda_i((A+B)^{-1}A)\}$ if

$$\max_{z \in R} \left| \prod_{n=0}^{N-1} (1 - \tau_n z) \right| = \min_{\{\hat{\tau}_n\}_{n=0}^{N-1}} \max_{z \in R} \left| \prod_{n=0}^{N-1} (1 - \hat{\tau}_n z) \right|.$$

More conveniently, we will say the best sequence in R .

In the remainder of this section a best sequence of parameters will be derived for a circular region and then for an elliptic region.

To begin the derivation the following definitions are required.

Definition 3.2: Let $g_n(z)$ be any complex polynomial of degree n , then $M[g_n(r)]$ denotes the maximum modulus function of $g_n(z)$ for $|z| \leq r$, i.e.,

$$M[g_n(r)] = \max_{|z| \leq r} |g_n(z)|.$$

Definition 3.3: S_n is the set of all polynomials g_n of degree n such that $g_n(1) = 1$.

Definition 3.4: $M^*[g_n(r, d)] = \max_{|d-z| \leq r} |g_n(z)|$.

Definition 3.5: S_n^* is the set of all polynomials g_n of degree n such that $g_n(0) = 1$.

The polynomials which are required in (3.14) belong to S_n^* . However we will need the following theorem, due to E.H. Zarantonello (see Varga [19]), which is stated for polynomials in S_n . Therefore, after stating the Theorem it will be modified in Theorem 3.2 to apply to polynomials in S_n^* .

Theorem 3.1: For all r such that $0 \leq r \leq 1$, $\min_{g_n \in S_n} \{M[g_n(r)]\} = r^n$ for all

positive integers n .

Theorem 3.2: For all real r and complex d such that $0 \leq r < |d|$,

$$\min_{g_n \in S_n^*} \{M^*[g_n(r, d)]\} = \left(\frac{|r|}{|d|} \right)^n$$

for all positive integers n .

Proof: Let r and d be such that $0 \leq r < |d|$ and suppose there exists some $g_n \in S_n^*$ such that $M^*[g_n(r, d)] < \left| \frac{r}{d} \right|^n$. Let $h_n(z) = g_n(d - dz)$. Then

$$\begin{aligned} \left| \frac{r}{d} \right|^n &> \max_{|d-z| \leq r} |g_n(z)| = \max_{|d-z| \leq r} \left| h_n\left(\frac{z-d}{-d}\right) \right| \\ &= \max_{|z| \leq \left| \frac{r}{d} \right|} |h_n(z)| = M\left[h_n\left(\left| \frac{r}{d} \right|\right)\right]. \end{aligned}$$

Since $h_n(1) = g_n(0) = 1$, the above inequality contradicts Theorem 3.1,

and $M^*[g_n(r, d)] \geq \left| \frac{r}{d} \right|^n$ for all g_n is S_n^* . On the other hand, if

$g_n(z) = \left(\frac{d-z}{d} \right)^n$ then $M^*[g_n(r, d)] = \left| \frac{r}{d} \right|^n$ and the conclusion of the theorem follows.

Corollary 3.2.1: The polynomial $P_n(z) = \left(\frac{d-z}{d} \right)^n$ is the polynomial in S_n^* satisfying $M^*[P_n(r, d)] = \min_{g_n \in S_n^*} M^*[g_n(r, d)]$ for all integers $n > 0$ and

all r and d such that $0 \leq r \leq |d|$.

Proof: This follows immediately from the proof of the theorem.

Corollary 3.2.1 can be used to determine a best sequence of τ_n 's to use on a disc $D = \{z: |z-d| < r, \text{ where } 0 < r < |d|\}$. That the

$\lambda_i((A+B)^{-1}A)$ lie in such a disc is seen as follows

The idea of choosing a sequence of τ_n 's is to speed convergence of iteration (2.5). It is reasonable to assume there is a constant value τ for which the iteration is convergent. From this and (3.3) it is seen that for some $\tau > 0$,

$$\lambda_i((A+B)^{-1}A) \in \delta\left(-\frac{1}{\tau}, 0; \frac{1}{\tau}\right) = \left\{z: \left|\frac{1}{\tau} - z\right| \leq \frac{1}{\tau}\right\}. \quad (3.15)$$

Consequently Corollary 3.2.1 can be used to determine a best sequence of τ_n 's in a disc as follows.

Theorem 3.3: If the eigenvalues of $(A+B)^{-1}A$ satisfy (3.15)

then the best sequence of τ_n 's to use in (2.5) is $\tau_n \equiv \tau$ a constant.

Moreover $\tau = \frac{1}{d}$ where d is the center of the smallest disc containing all $\lambda_i((A+B)^{-1}A)$.

Proof: Since no confusion will arise, λ_i will be used to denote $\lambda_i((A+B)^{-1}A)$ in this proof. Let $\delta(d, 0; r)$ be the smallest disc containing all the λ_i . That the smallest disc is of this form is seen as follows. Since the complex eigenvalues occur in complex conjugate pairs, the smallest disc containing them has a real center, $(d, 0)$. In addition (3.15) implies $0 \leq r < d$. Now from Theorem 3.2 and Corollary 3.2.1 it is seen that of all polynomials in S_n^* , the minimax polynomial on $\delta(d, 0; r)$ is

$$P_N(z) = \left(\frac{d-z}{d}\right)^N$$

$$\text{and } M^*[P_N(r, d)] = \left|\frac{r}{d}\right|^N.$$

It is apparent that either at least 4 of the λ_i lie on the

circumference as in Figure 3.1a, at least 2 as in Figure 3.1b or 3 with one on the real axis and two as in Figure 2.1a.

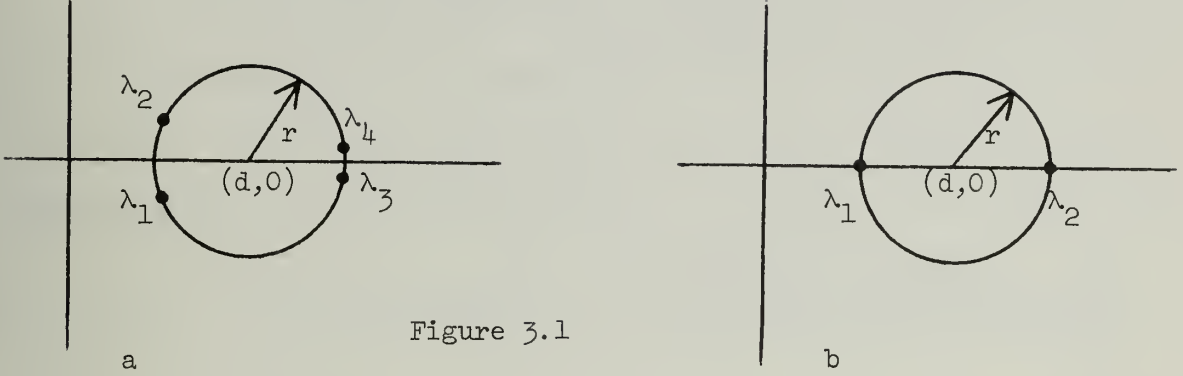


Figure 3.1

Let $\delta(d+\epsilon_1, \epsilon_2; r')$ be any other disc containing all the λ_i . The polynomial $g_N \in S_N^*$ which is the minimax polynomial on this disc satisfies

$$M^*[q_N(r', (d+\epsilon_1, \epsilon_2))] = \left[\frac{r'}{[(d+\epsilon_1)^2 + (\epsilon_2)^2] \frac{1}{2}} \right]^N. \quad (3.17)$$

The Theorem will be established if it is shown that the quantity in (3.17) is greater than the quantity in (3.16). To establish this we shall assume, without loss of generality, that ϵ_1 and ϵ_2 are non-negative. If this were not the case, the following argument would be changed only in the selection of the λ_i of Figure 3.1.

Since $\delta(d+\epsilon_1, \epsilon_2; r')$ contains all the λ_i it in particular contains one which is as λ_1 of Figure 3.1a or Figure 3.1b, i.e., $\lambda_1 = (u, -v)$ where $u > 0$ and $v \geq 0$. Therefore,

$$\begin{aligned}
(r')^2 &\geq (d+\epsilon_1-u)^2 + (\epsilon_2+v)^2 \\
&= (d+\epsilon_1)^2 + u^2 + \epsilon_2^2 + v^2 + 2\epsilon_2 v - 2u(d+\epsilon_1) \\
&\quad (d+\epsilon_1)^2 + u^2 + \epsilon_2^2 + v^2 - 2u[(d+\epsilon_1)^2 + \epsilon_2^2]^{\frac{1}{2}} \\
&= [((d+\epsilon_1)^2 + \epsilon_2^2)^{\frac{1}{2}} - u]^2 + v^2.
\end{aligned}$$

Hence the quantity in (3.17) is greater than

$$\left[\frac{\{[(d+\epsilon_1)^2 + \epsilon_2^2]^{\frac{1}{2}} - u\}^2 + v^2}{(d+\epsilon_1)^2 + \epsilon_2^2} \right]^{\frac{N}{2}}.$$

Thus

$$M^*[q_N(r', (d+\epsilon_1+\epsilon_2))] > \left[\frac{(d+\epsilon_3-u)^2 + v^2}{(d+\epsilon_3)^2} \right]^{\frac{N}{2}},$$

where $\epsilon_3 = [(d+\epsilon_1)^2 + \epsilon_2^2]^{\frac{1}{2}} - d > 0$.

And

$$M^*[q_N(r', (d+\epsilon_1, \epsilon_2))] > \left[\frac{(d-u)^2 + v^2}{d^2} \right]^{\frac{N}{2}} = \left| \frac{r}{d} \right|^N = M^*[P_N(d, r)].$$

The last inequality is derived from $d > u$ and the Law of Sines.

We have proved that the polynomial $P_N(z)$ of (3.16) has the minimum maximum on the smallest disc containing each λ_i , where the minimum is taken over all polynomials in S_N^* ; and this maximum is less than the minimum maximum of all polynomials in S_N^* on any other disc containing the λ_i . The τ_n should, therefore be chosen so that

$$\prod_{n=0}^{N-1} (1 - \tau_n z) = \left(\frac{d-z}{d} \right)^N = \left(1 - \frac{1}{d} z \right)^N,$$

i.e., $\tau_n \equiv \frac{1}{d}$. This completes the proof.

If more is known about the $\lambda_i((A+B)^{-1}A)$ then some more useful results concerning a best sequence of parameters can be obtained. In particular, if the $\lambda_i((A+B)^{-1}A)$ are contained in an ellipse then the best sequence is a Tchebychev sequence as in the real case. Before proving this we need to recall some terminology

Definition 3.6: In the ellipse $\frac{(x-p)^2}{a^2} + \frac{(y-q)^2}{b^2} = 1$ with $a^2 \geq b^2$, a is called the semi-major axis, $e = (1 - \frac{b^2}{a^2})^{\frac{1}{2}}$ is the eccentricity and (p,q) is the center.

A modification of the following result due to Clayton, see Wrigley [23], will be used in determining a best sequence of τ_n 's when the $\lambda_i((A+B)^{-1}A)$ are contained in an ellipse.

Theorem 3.4: The polynomial in S_N which, over the ellipse of semi-major axis $a < 1$, eccentricity e and center at the origin, has the smallest maximum modulus value is given by

$$P_N(z) = T_N(z/ae) / T_N(1/ae) . \quad (3.18)$$

If the center of the ellipse, still assumed to lie within the unit circle, is translated to $(d,0)$ then the minimax polynomial is

$$P_N(z) = T_N((z-d)/ae) / T_N((1-d)/ae) . \quad (3.19)$$

Remark: As $e \rightarrow 0$ the ellipse becomes a circle and, as expected, the

polynomial P_N of (3.18) approaches z^N . This is easily seen, for when e is small,

$$\begin{aligned} T_N(z/ae) &= \frac{1}{2} [\exp(N \cosh^{-1}(z/ae)) + \exp(-N \cosh^{-1}(z/ae))] \\ &\cong \frac{1}{2} \exp(N \cosh^{-1}(z/ae)) \\ &= \frac{1}{2} \left[\frac{z}{ae} + ((z/ae)^2 - 1)^{\frac{1}{2}} \right]^N \\ &\cong 2^{N-1} (z/ae)^N. \end{aligned}$$

Hence $\lim_{e \rightarrow 0} P_N(z) = \frac{2^{N-1} (z/ae)^N}{2^{N-1} (1/ae)^N} = z^N$. In the first equality the fact that $T_N(z) = \cosh(N \cosh^{-1} z)$ is valid throughout the complex plane was used (see for example Wrigley [23]).

Since we are interested in polynomials in S_N^* rather than S_N , Theorem 3.4 is rephrased as follows.

Theorem 3.5: Let E be the ellipse with semi-major axis a , eccentricity e and center $(d, 0)$ where $a < d$. The polynomial in S_N^* which has the smallest maximum modulus value on E is given by

$$P_N(z) = \frac{T_N\left(\frac{d-z}{ae}\right)}{T_N(d/ae)}. \quad (3.20)$$

Proof: First note that $P_N(z)$ of (3.20) is in S_N^* . Now suppose there is some $q_N(z)$ in S_N^* with smaller maximum modulus than P_N on E . Then

$$\max_{z \in E'} \left| q_N(d-dz) \right| < \max_{z \in E'} \left| P_N(d-dz) \right|, \quad (3.21)$$

where E' is the ellipse with semi-major axis a/d , eccentricity e and center $(0,0)$. But

$$P_N(d-dz) = \frac{T_N\left(\frac{z}{ea/d}\right)}{T_N\left(\frac{1}{e \frac{a}{d}}\right)}.$$

Thus (3.21) contradicts Theorem 3.4 and the conclusion of the theorem follows.

The following corollary shows that polynomial $P_N(z)$ in (3.20) can be put in the form of the polynomial (3.12). Thus the best sequence of τ_n 's in an ellipse can be implemented in the same way that the best sequence on an interval (e_1, e_2) is implemented.

Corollary 3.5.1: Let E be as in Theorem 3.5 and suppose each $\lambda_1((A+B)^{-1}A)$ is in E . The best sequence of parameters to use in (2.5) on E is the sequence $\{\tau_n\}$ for which

$$\prod_{n=0}^{N-1} (1-\tau_n z) = P_N(z) = \frac{T_N((e_1+e_2-2z)/(e_2-e_1))}{T_N((e_1+e_2)/(e_2-e_1))} \quad (3.22)$$

where $e_1 = d-ae$ and $e_2 = d+ae$.

Proof: Let E be as in Theorem 3.5. As noted when considering a disc, the existence of such an E which contains each λ_1 is guaranteed by (3.15). Let $e_1 = d-ae$ and $e_2 = d+ae$. The best sequence of parameters on E yields a polynomial

$$P_N(z) = \prod_{n=0}^{N-1} (1 - \tau_n z) = \frac{T_N\left(\frac{d-z}{ae}\right)}{T_N\left(\frac{d}{ae}\right)} \quad (3.23)$$

But $e_1 + e_2 = 2d$ and $e_2 - e_1 = 2ae$ so that (3.23) is

$$P_N(z) = \frac{T_N\left(\frac{\frac{1}{2}(e_1 + e_2) - z}{\frac{1}{2}(e_2 - e_1)}\right)}{T_N\left(\frac{e_2 + e_1}{e_2 - e_1}\right)} = \frac{T_N\left(\frac{e_2 + e_1 - 2z}{e_2 - e_1}\right)}{T_N\left(\frac{e_2 + e_1}{e_2 - e_1}\right)}$$

as claimed.

Remark: As the ellipse degenerates to a line, $e \rightarrow 1$ so that

$$e_1 \rightarrow d-a \quad \text{and} \quad e_2 \rightarrow d+a$$

which is exactly the result of the previous section.

4. USING A SEQUENCE OF τ 's WHEN LITTLE IS KNOWN ABOUT THE EIGENVALUES OF $(A+B)^{-1}A$

4.1 Introduction

In this chapter we will develop an always convergent algorithm to solve $Ax=q$. The algorithm is similar to the method presented in section 3.3 but is superior since it does not require any knowledge of the eigenvalues of $(A+B)^{-1}A$. The theoretical basis is derived in sections 4.2, 4.3 and 4.4. The algorithm is stated in 4.5 and some experimental results are given in 4.6.

The sequence of τ_n 's in the iteration

$$(A+B)x_{n+1} = (A+B)x_n - \tau_n(Ax_n - q) \quad (4.1)$$

is defined during the execution of the algorithm. As in the methods presented in Chapter 3, the sequence is dependent upon an interval. Instead of using a constant interval as the previous algorithms do; the algorithm of this chapter generates a sequence of intervals (a_i, b_i) which converge to the optimal interval.

The following is a brief description of the algorithm. It is shown that $\delta_n = x_n - x_{n-1}$, which is automatically calculated in the practical implementation of (4.1), can be made to approach an eigenvector of $(A+B)^{-1}A$ corresponding to an extreme eigenvalue. This approximate eigenvector is used to compute an approximation to the corresponding eigenvalue. The interval (a_i, b_i) is then improved by replacing one endpoint with the approximate eigenvalue.

The algorithm to be developed can be applied to any iteration of the form (4.1) in which A and $A+B$ are positive definite. If A is derived from the discretization of an elliptic partial differential equation as described in Chapter 2 and $A+B$ is defined by one of the factorizations due to Kim [8], Stone [15], or Dupont, Kendall and Rachford [5] then A and $A+B$ will be positive definite. Thus the algorithm can be applied to these factorizations.

It is assumed in the sequel that A and $A+B$ are positive definite.

Additional Notation: Throughout this chapter the subscripts and superscripts will denote the following:

- i corresponds to the interval (a_i, b_i) .
- n and N correspond to the number of iterations.
- k and K correspond to the dimension of the matrix A .

Thus the vector v calculated on the n^{th} iteration using the interval (a_i, b_i) will be denoted by v_n^i . This vector will be denoted by v_n when the interval used is not being stressed.

The algorithm will use vectors which are calculated during execution of computer programs written to solve the set of equations $Ax=q$ by using a factorization and iteration (4.1). Therefore, we shall begin with a description of the practical utilization of the methods presented in Chapter 3.

Two methods are presented in Section 3.3. In the first a vector x_0 , an interval (a,b) and an integer N are selected. The iterates x_n , $n=1, \dots, N$, are then defined by

$$(A+B)x_{n+1} = (A+B)x_n - \tau_n(Ax_n - q) \quad (4.1)$$

where the τ_n are selected so that

$$P_N(X) = \prod_{n=0}^{N-1} (1 - \tau_n X) = \frac{T_N\left(\frac{a+b-2X}{b-a}\right)}{T_N\left(\frac{a+b}{b-a}\right)}.$$

In the second method the iterates are defined by

$$\begin{aligned}\hat{x}_{n+1} &= \hat{x}_n + \delta \hat{x}_n, \\ \delta x_n &= \frac{4T_n(y)}{(b-a)T_{n+1}(y)}(A+B)^{-1}(q - A\hat{x}_n) + \frac{T_{n-1}(y)}{T_{n+1}(y)}\delta \hat{x}_{n-1}, \\ \delta \hat{x}_0 &= \frac{2}{a+b}(A+B)^{-1}(q - A\hat{x}_0),\end{aligned}\tag{4.2}$$

where $y = \frac{(a+b)}{(b-a)}$ and $\hat{x}_0$ is the initial vector.

Throughout the remainder of this chapter x_n will denote a vector defined by (4.1) and $\hat{x}_n$ will denote a vector defined by (4.2).

The vectors in the following definitions are calculated in the practical use of (4.1).

Definition 4.1: The n^{th} residual is the vector $r_n = q - Ax_n$.

Definition 4.2: The n^{th} incremental vector is $\delta_n = x_n - x_{n-1}$.

Using these definitions the iteration

$$(A+B)x_{n+1} = (A+B)x_n - \tau_n(Ax_n - q)$$

can be rewritten as

$$(A+B)\delta_n = \tau_{n-1}r_{n-1}.$$

This equation is solved by replacing $A+B$ by LU and calculating δ_n from

$$Lz = \tau_{n-1}r_{n-1}$$

and

$$U\delta_n = z.$$

The next iterate is then $x_n = x_{n-1} + \delta_n$.

The vectors $\hat{x}_n$ are similarly calculated. Instead of δ_n and r_n however, we define

Definition 4.3: $\hat{r}_n = q - A\hat{x}_n$.

Iteration (4.2) is performed by computing

$$z = L^{-1}\hat{r}_n,$$

$$\hat{\delta}_{n+1} = U^{-1}z,$$

and

$$\delta\hat{x}_n = \frac{4T_n(y)}{(b-a)T_{n+1}(y)} \hat{\delta}_{n+1} + \frac{T_{n-1}(y)}{T_{n+1}(y)} \delta\hat{x}_{n-1} \quad (4.3)$$

$$n \geq 1.$$

The vector $\delta\hat{x}_0$ is computed from

$$z = L^{-1}\hat{r}_0$$

and

$$\delta\hat{x}_0 = \frac{2}{a+b} U^{-1}z.$$

For every $n \geq 0$ x_{n+1} is calculated from

$$\hat{x}_{n+1} = \hat{x}_n + \delta\hat{x}_n.$$

Note that δ_n defined in (4.3) satisfies

$$\hat{\delta}_n = (A+B)^{-1}\hat{r}_{n-1} \quad (4.4)$$

The importance of this vector will be shown in Section 4.4.

4.2 Calculating An Eigenvector And Eigenvalue Of $(A+B)^{-1}A$

In this section it is proved that if y_0 is any vector and if

$$y_N = P_N((A+B)^{-1}A)y_0$$

where

$$P_N(X) = \frac{T_N\left(\frac{a+b-2X}{b-a}\right)}{T_N\left(\frac{a+b}{b-a}\right)} \quad (4.5)$$

and $a > \lambda_{\min}$ or $b < \lambda_{\max}$ then y_N approaches an eigenvector of $(A+B)^{-1}A$.

It will also be proved that

$$\mu_N = \frac{\langle Ay_N, y_N \rangle}{\langle (A+B)y_N, y_N \rangle} \quad (4.6)$$

approaches the corresponding eigenvalue.

In the next section δ_{N+1} is shown to satisfy $\delta_{N+1} = P_N((A+B)^{-1}A)\delta_1$. From this it follows that δ_{N+1} is approaching an eigenvector of $(A+B)^{-1}A$ as N increases.

We shall begin with some motivation.

Suppose that y_0 is any initial vector and that

$$y_N = \prod_{n=0}^{N-1} (I - \tau_n^*(A+B)^{-1}A)y_0 = P_N^*((A+B)^{-1}A)y_0, \text{ where } P_N^*(X) = \prod_{n=0}^{N-1} (I - \tau_n^*X). \quad (4.7)$$

The object is to select the τ_n^* such that y_N approaches one of the extreme eigenvalues of $(A+B)^{-1}A$. If

$$y_0 = \sum_{k=1}^K \alpha_k w_k, \text{ where the } w_k, k=1, \dots, K, \text{ form a complete set}$$

of eigenvectors of $(A+B)^{-1}A$, then y_N is given by

$$\begin{aligned} y_N &= \sum_{k=1}^K \alpha_k P_N^* \left[\lambda_k((A+B)^{-1}A) \right] w_k \\ &= P_N^* \left[\lambda_1((A+B)^{-1}A) \right] \sum_{k=1}^K \alpha_k \frac{P_N^* \left[\lambda_k((A+B)^{-1}A) \right]}{P_N^* \left[\lambda_1((A+B)^{-1}A) \right]} w_k. \end{aligned}$$

To make y_N most rapidly approach w_1 , $\{\tau_n^*\}$ in (4.7) should be selected so that

$$\text{Max}_{k=2, \dots, K} \left| \frac{P_N^* \left[\lambda_k((A+B)^{-1}A) \right]}{P_N^* \left[\lambda_1((A+B)^{-1}A) \right]} \right|$$

is minimized. Since $\lambda_2, \dots, \lambda_K$ are unknown, we minimize

$$\max_{\lambda_K \leq X \leq \lambda_2} \frac{P_N^*(X)}{P_N^*[\lambda_1((A+B)^{-1}A)]} \quad (4.8)$$

where λ_K denotes $\lambda_K((A+B)^{-1}A)$.

Note that y_n converges to an eigenvector of $(A+B)^{-1}A$ provided

$$\lim_{N \rightarrow \infty} \max_{k=2, \dots, K} \frac{\left| P_N^*[\lambda_k((A+B)^{-1}A)] \right|}{\left| P_N^*[\lambda_1((A+B)^{-1}A)] \right|} = 0, \quad (4.9)$$

or

$$\lim_{N \rightarrow \infty} \max_{\lambda_K \leq X \leq \lambda_2} \frac{\left| P_N^*(X) \right|}{\left| P_N^*(\lambda_1) \right|} = 0.$$

Minimizing (4.8) merely speeds the convergence.

The polynomial P_N^* , which minimizes (4.8), is a scaled and translated Tchebychev polynomial as seen by the corollary to the following theorem.

Theorem 4.1: Let $M(P) = \max_{-1 \leq X \leq 1} |P(X)|$. The Tchebychev polynomial $T_N(X)$ of degree N has the property that

$$\frac{M(T_N)}{|T_N(y)|} \leq \frac{M(q_N)}{|q_N(y)|}$$

for all N degree polynomials q_N and all $y \notin [-1, 1]$.

Proof: The theorem is a restatement of the classical result of Tchebychev that $M(T_N) \leq M(q_N)$ for all q_N such that $q_N(1) = 1$, and $|T_N(y)| \geq |q_N(y)|$ for all $y \notin [-1, 1]$.

Corollary 4.1.1: The polynomial

$$P_N^*(X) = \frac{T_N\left(\frac{a+b-2X}{b+a}\right)}{T_N\left(\frac{a+b}{b-a}\right)}$$

satisfies

$$\frac{\max_{a \leq X \leq b} |P_N^*(X)|}{|P_N^*(y)|} \leq \frac{\max_{a \leq X \leq b} |q_N(X)|}{|q_N(y)|}$$

for each N degree polynomial q_N such that $q_N(0)=1$ and for all $y \notin [a,b]$.

Proof: Omitted.

Theorem 4.2: Let w_k , $k=1, \dots, K$, form a complete set of eigenvectors of $(A+B)^{-1}A$ corresponding to the positive eigenvalues

$$\lambda_{MAX} = \lambda_1 > \lambda_2 \geq \dots \geq \lambda_{K-1} > \lambda_K = \lambda_{min} > 0.$$

Let $y_0 = \sum_{k=1}^K \alpha_k w_k$ be any vector such that $\alpha_1 \alpha_K \neq 0$ and define the sequence of τ_n , $n=0, \dots, N-1$, such that

$$P_N(X) = \prod_{n=0}^{N-1} (1 - \tau_n X) = \frac{T_N\left(\frac{a+b-2X}{b-a}\right)}{T_N\left(\frac{a+b}{b-a}\right)}, \quad (4.10)$$

where $a > \lambda_{min}$ or $b < \lambda_{MAX}$ and $|P_N(\lambda_{min})| \neq |P_N(\lambda_{MAX})|$. If $y_N = P_N((A+B)^{-1}A)y_0$ then either

$$\lim_{N \rightarrow \infty} \frac{1}{P_N(\lambda_{MAX})} y_N = \alpha_1 w_1$$

or

(4.11)

$$\lim_{N \rightarrow \infty} \frac{1}{P_N(\lambda_{min})} y_N = \alpha_K w_K.$$

The convergence of y_N to w_1 or w_K is most rapid when $a = \lambda_{\min}$ and $b = \lambda_2$ or when $a = \lambda_{k-1}$ and $b = \lambda_{\max}$, respectively. And y_N converges to w_1 only if $b < \lambda_{\max}$ and converges to w_K only if $a > \lambda_{\min}$.

Proof: Suppose $a > \lambda_{\min}$ or $b < \lambda_{\max}$.

Suppose $\left| P_N(\lambda_{\min}) \right| < \left| P_N(\lambda_{\max}) \right|$. This implies $b < \lambda_{\max}$ and

$$\left| P_N(\lambda_{\max}) \right| > \left| P_N(\lambda_k) \right| \quad k = 1, \dots, K. \quad (4.12)$$

From the definition of y_0 and y_N we have

$$\begin{aligned} \frac{1}{P_N(\lambda_{\max})} y_N &= \sum_{k=1}^K \alpha_k \frac{P_N(\lambda_k)}{P_N(\lambda_{\max})} w_k \\ &= \sum_{k=1}^K \alpha_k \frac{T_N\left(\frac{a+b-2\lambda_k}{b-a}\right)}{T_N\left(\frac{a+b-2\lambda_{\max}}{b-a}\right)} w_k \\ &= \alpha_1 w_1 + \sum_{\lambda_k \in [a, b]} \alpha_k \frac{T_N\left(\frac{a+b-2\lambda_k}{b-a}\right)}{T_N\left(\frac{a+b-2\lambda_{\max}}{b-a}\right)} w_k \\ &\quad + \sum_{\substack{\lambda_k \notin [a, b] \\ \lambda_k \neq \lambda_{\max}}} \alpha_k \frac{T_N\left(\frac{a+b-2\lambda_k}{b-a}\right)}{T_N\left(\frac{a+b-2\lambda_{\max}}{b-a}\right)} w_k \end{aligned}$$

Applying Corollary 4.1.1 to the terms of the first sum we see that as N increases these terms are decreasing at an optimal rate.

For the terms of the second sum we have from (4.12) that

$$1 < \frac{|a+b-2\lambda_k|}{|b-a|} < \frac{|a+b-2\lambda_{\max}|}{|b-a|}$$

Thus

$$\cosh^{-1} \left| \frac{a+b-2\lambda_k}{b-a} \right| - \cosh^{-1} \left| \frac{a+b-2\lambda_{\text{MAX}}}{b-a} \right| < 0.$$

Using the inequalities and the relation

$$T_N(X) = [\exp(N \cosh^{-1} X) + \exp(-N \cosh^{-1} X)]/2$$

for $|X| > 1$ yields

$$\lim_{N \rightarrow \infty} \frac{T_N \left(\frac{a+b-2\lambda_k}{b-a} \right)}{T_N \left(\frac{a+b-2\lambda_{\text{MAX}}}{b-a} \right)} = 0.$$

Thus $\lim_{N \rightarrow \infty} \frac{1}{P_N(\lambda_{\text{MAX}})} y_N = \alpha_1 w_1$ which establishes the first half of (4.11).

The second half of (4.11) is similarly derived assuming $|P_N(\lambda_{\text{min}})| > |P_N(\lambda_{\text{MAX}})|$.

That the convergence is most rapid when $a = \lambda_{\text{min}}$ and $b = \lambda_2$ or

$a = \lambda_{K-1}$ and $b = \lambda_{\text{MAX}}$ follows directly from Corollary 4.1.1.

Finally we note that $y_1, y_2, \dots$ converges to w_1 only if

$|P_N(\lambda_{\text{MAX}})| > |P_N(\lambda_{\text{min}})|$, i.e., only if $b < \lambda_{\text{MAX}}$. Also $y_1, y_2, \dots$ converges to w_K only if $a > \lambda_{\text{min}}$. This completes the proof.

Any approximation to an eigenvector of $(A+B)^{-1}A$ yields a

corresponding eigenvalue approximation. This is derived from a theorem in [20], presented here for convenience.

Theorem 4.3 [20]: If A and B are positive definite then the eigenvectors, x_k , and the corresponding eigenvalues, λ_k , of the generalized eigenvalue problem,

$$Ax_k = \lambda_k Bx_k,$$

satisfy the following:

1, $\lambda_k > 0$;

2, $x_k^T Bx_j = 0 \quad j \neq k$.

Proof: Since B is positive definite there exists an orthogonal R such that $R^T B R = \text{diag}(\beta_k^2) = D^2$, where β_k is an eigenvalue of B.

$$(A - \lambda B) = R D (D^{-1} R^T A R D^{-1} - \lambda I) D R^T.$$

Hence $\det(A - \lambda B) = (\det R)^2 (\det D)^2 \det(P - \lambda I)$, where $P = D^{-1} R^T A R D^{-1}$. Since $(\det R)^2 = 1$ and $(\det D) = \Pi(\beta_k^2) > 0$, the zeros of $\det(A - \lambda B)$ are those of $\det(P - \lambda I)$. These are the eigenvalues of a symmetric matrix and are therefore real. The matrix P has a complete set of eigenvectors z_k which may be taken to be orthogonal. Hence

$D^{-1} R^T A R D^{-1} z_k = \lambda_k z_k$; and $A(R D^{-1} z_k) = \lambda_k R D z_k = \lambda_k R D (D R^T R D^{-1}) z_k = \lambda_k B(R D^{-1} z_k)$. Thus $x_k = R D^{-1} z_k$ is an eigenvector of the generalized eigenvalue problem, corresponding to λ_k .

Since $\{z_k\}$ is orthogonal and $R D^{-1}$ is nonsingular, $\{x_k\}$ forms a complete set of real eigenvectors. Furthermore

$$0 = z_k^T z_j = (D R^T x_k)^T D R^T x_j = x_k^T R D D R^T x_j = x_k^T B x_j,$$

so that $\{x_k\}$ is orthogonal with respect to B.

Finally, $\lambda_k > 0$ since $\langle A x_k, x_k \rangle = \langle \lambda_k B x_k, x_k \rangle$. This completes the proof.

Observe that this theorem applied to A and A+B implies that the eigenvectors, w_k , of $(A+B)^{-1} A$ are orthogonal with respect to (A+B).

The next lemma defines an approximate eigenvalue.

Lemma 4.1: Suppose A+B and A are positive definite. Let $y = \sum_{k=1}^K \epsilon_k w_k$

where w_k is an eigenvector of $(A+B)^{-1} A$ corresponding to λ_k . Let

$\mu = \frac{\langle y, A y \rangle}{\langle y, (A+B) y \rangle}$. If y is approximately equal to w_j with errors ϵ_k in the

direction w_k , and $\epsilon_J \gg \epsilon_k$ $k \neq J$, then $\mu \cong \lambda_J$ with error of order $\left(\frac{\epsilon_k}{\epsilon_J}\right)^2$.

Proof: Assume $J = 1$. The remaining cases are similar. Assume $\{w_k\}$ is orthonormal with respect to $(A+B)$. Let

$$\mu = \frac{\langle y, Ay \rangle}{\langle y, (A+B)y \rangle} \quad (4.13)$$

Substituting $\sum_{k=1}^K \epsilon_k w_k$ for y and $(A+B)(A+B)^{-1}A$ for A in (4.13) yields

$$\mu = \frac{\sum_{k=1}^K \epsilon_k w_k^T (A+B)(A+B)^{-1} A \sum_{j=1}^K \epsilon_j w_j}{\sum_{k=1}^K \epsilon_k w_k^T (A+B) \sum_{j=1}^K \epsilon_j w_j}.$$

Now using the facts that $(A+B)^{-1}Aw_j = \lambda_j w_j$ and $w_k^T (A+B)w_j = \delta_{j,k}$, where $\delta_{j,k}$ is Kronecker delta, we obtain

$$\mu = \frac{\sum_{k=1}^K \epsilon_k^2 \lambda_k}{\sum_{k=1}^K \epsilon_k^2} = \frac{\lambda_1 + \sum_{k=2}^K \epsilon_k^2 \lambda_k / \epsilon_1^2}{1 + \sum_{k=2}^K \epsilon_k^2 / \epsilon_1^2}.$$

This completes the proof.

In the case $A+B=I$ the above lemma reduces to a result of Hestenes [21].

4.3 Application Of The Method

To compute an eigenvalue using the above method it is necessary to determine a sequence of vectors $y_0, \dots, y_N$ such that $y_N = P((A+B)^{-1}A)y_0$ where

$$P_N(X) = T_N \left[\frac{a+b-2X}{b-a} \right] / T_N \left[\frac{a+b}{b-a} \right].$$

The next theorem shows that δ_n , $n = 1, \dots, N+1$ is the desired sequence of vectors. Since δ_n is calculated when iteration (4.1) is used to solve $Ax = q$, no additional work is required to obtain the approximate eigenvector.

Theorem 4.4: Let x_0 be any vector and define x_n , $n = 1, 2, \dots, N+1$ by iteration (4.1). If the sequence of $\{\tau_n\}$ used in the iteration is such that

$$P_N(X) = \prod_{n=0}^{N-1} (1 - \tau_n X) = \frac{T_N\left(\frac{a+b-2X}{b-a}\right)}{T_N\left(\frac{a+b}{b-a}\right)},$$

where $a > \lambda_{\min}$ or $b < \lambda_{\max}$ and $\left|P_N(\lambda_{\min})\right| \neq \left|P_N(\lambda_{\max})\right|$, then $\delta_{N+1} = x_{N+1} - x_N$ approaches an eigenvector of $(A+B)^{-1}A$ as N increases.

Proof: Since $\delta_n = x_n - x_{n-1} = (I - \tau_{n-1}(A+B)^{-1}A)x_{n-1} + \tau_{n-1}(A+B)^{-1}q - x_{n-1}$,

$$\delta_n - \tau_{n-1}(A+B)^{-1}q = -\tau_{n-1}(A+B)^{-1}Ax_{n-1} \quad (4.14)$$

and

$$\delta_{n+1} - \tau_n(A+B)^{-1}q = -\tau_n(A+B)^{-1}Ax_n. \quad (4.15)$$

Subtracting (τ_{n-1}/τ_n) times (4.15) from (4.14) yields

$$\delta_n - (\tau_{n-1}/\tau_n)\delta_{n+1} = \tau_{n-1}(A+B)^{-1}A(x_n - x_{n-1}) = \tau_{n-1}(A+B)^{-1}A\delta_n,$$

which implies $(I - \tau_{n-1}(A+B)^{-1}A)\delta_n = (\tau_{n-1}/\tau_n)\delta_{n+1}$.

Hence $(\tau_n/\tau_{n-1})(I - \tau_{n-1}(A+B)^{-1}A)\delta_n = \delta_{n+1}$. And

$$\tau_N/\tau_0 \prod_{n=0}^{N-1} (I - \tau_n(A+B)^{-1}A)\delta_1 = \delta_{N+1}.$$

Since the sequence of τ_n , $n = 0, \dots, N-1$, is used cyclically, $\tau_N = \tau_0$

and thus

$$\delta_{N+1} = \prod_{n=0}^{N-1} (I - \tau_n (A+B)^{-1} A) \delta_1. \quad (4.16)$$

The conclusion follows from Theorem 4.2.

An approximation to the corresponding eigenvalue can be calculated, according to Lemma 4.1, by

$$\mu_N = \frac{\langle \delta_{N+1},^A \delta_{N+1} \rangle}{\langle \delta_{N+1},^{(A+B)} \delta_{N+1} \rangle}. \quad (4.17)$$

4.4 Using The Recursive Property Of The Tchebychev Polynomials

From the theory developed in Sections 4.2 and 4.3 it is possible to use iteration (4.1) to solve $Ax=q$ and simultaneously compute an approximate eigenvector of $(A+B)^{-1}A$. The process is started by selecting N and using (4.1) to compute the sequence of vectors x_n , $n=1, \dots, N+1$. The resulting approximate eigenvector is $\delta_{N+1} = x_{N+1} - x_N$. In the next theorem it is shown that if $\hat{x}_n$ is computed from (4.2) then

$$\hat{\delta}_{n+1} = (A+B)^{-1}(q - A\hat{x}_n) \quad (4.18)$$

is an approximate eigenvector for every $n=1, 2, \dots$. Recall that

$\hat{\delta}_{n+1} \neq \hat{x}_{n+1} - \hat{x}_n$. Note that $\hat{\delta}_{n+1}$ is computed when (4.2) is used to solve

$Ax=q$ and thus the approximate eigenvector is again obtained with no additional work.

Theorem 4.5: Let $\hat{x}_0$ be any vector and define the sequence of vectors

$\hat{x}_n$, $n=1, 2, \dots$, by iteration (4.2). Assume the interval (a,b) used in the

iteration satisfies $(a, b) \not\subset (\lambda_{\min}, \lambda_{\max})$. Let $\hat{r}_n = q - A\hat{x}_n$. It follows that $\hat{\delta}_n = (A+B)^{-1}\hat{r}_n$ approaches an eigenvector of $(A+B)^{-1}A$ as n increases. Furthermore $\hat{\delta}_n$ approaches the eigenvector corresponding to $\lambda_{\max}$ only if $b < \lambda_{\max}$ and approaches the eigenvector corresponding to $\lambda_{\min}$ only if $a > \lambda_{\min}$.

Proof: Define τ_n^N $n = 0, \dots, N-1$, such that

$$\prod_{n=0}^{N-1} (1 - \tau_n^N x) = \frac{T_N\left[\frac{a+b-2x}{b-a}\right]}{T_N\left[\frac{a+b}{b-a}\right]},$$

and let $\tau_n^N = \tau_{n \pmod N}^N$ for $n \geq N$. Let $x_0^N = x_0$ and define

$$x_n^N = x_{n-1}^N + \delta_n^N \quad \text{where} \quad \delta_n^N = \tau_{n-1}^N (A+B)^{-1} (q - Ax_{n-1}^N).$$

The vectors $\hat{x}_n$ defined by (4.2) have the property that

$$\hat{x}_n = x_n^N \quad \text{for every } n.$$

$$\text{Therefore } \hat{\delta}_{n+1} = (A+B)^{-1} (q - A\hat{x}_n) = (A+B)^{-1} (q - Ax_n^N) = \frac{\tau_n^N}{\tau_n^N} (A+B)^{-1} (q - Ax_n^N).$$

But $\tau_n^N = \tau_0^N$ and $\tau_n^N (A+B)^{-1} (q - Ax_n^N) = \delta_{n+1}^N$. Hence

$$\hat{\delta}_{n+1} = \frac{1}{\tau_0^N} \delta_{n+1}^N.$$

From (4.16) we obtain

$$\hat{\delta}_{n+1} = \frac{1}{\tau_0^N} \prod_{j=0}^{n-1} (I - \tau_j^N (A+B)^{-1} A) \delta_1^N. \quad (4.19)$$

$$\text{Since } \frac{1}{\tau_0^N} \delta_1^N = \frac{1}{\tau_0^N} \left[\tau_0^N (A+B)^{-1} (q - Ax_0^N) \right] = (A+B)^{-1} (q - A\hat{x}_0) = \hat{\delta}_1,$$

(4.19) can be rewritten as

$$\hat{\delta}_{n+1} = \prod_{j=0}^{n-1} (I - \tau_j^n (A+B)^{-1} A) \hat{\delta}_1. \quad (4.20)$$

The conclusion follows from Theorem 4.2.

The approximate eigenvalue is now given by

$$\mu_n = \frac{\langle \hat{\delta}_{n+1}, A \hat{\delta}_{n+1} \rangle}{\langle \hat{\delta}_{n+1}, (A+B) \hat{\delta}_{n+1} \rangle}. \quad (4.21)$$

We now show that no matrix-vector multiplications are necessary to compute the quantities $(A+B)\hat{\delta}_{n+1}$ and $A\hat{\delta}_{n+1}$ which are needed in the calculation of μ_n . The vector $(A+B)\hat{\delta}_{n+1} = \hat{r}_n$ and therefore is available. That $A\hat{\delta}_{n+1}$ can be calculated without a matrix-vector multiplication is proved in the next theorem. More precisely, it is shown that $A\hat{\delta}_{n+1}$ may be calculated from known quantities with only two vector additions, at the expense of additional storage to hold $(\hat{r}_{n-1} - \hat{r}_n)$.

Theorem 4.6: If $\hat{x}_n$ is the n^{th} iterate defined by (4.2) and $\hat{r}_n = q - A\hat{x}_n$, then

$$A\hat{\delta}_{n+1} = \frac{T_{n+2}(y)(b-a)}{4T_{n+1}(y)} \left[(\hat{r}_{n+1} - \hat{r}_{n+2}) - q_{n+1}(\hat{r}_n - \hat{r}_{n+1}) \right],$$

where $y = (a+b)/(b-a)$ as in (4.2) and $q_{n+1} = (T_n(y))/(T_{n+2}(y))$.

Proof: From (4.2) we have $\hat{x}_{n+2} = \hat{x}_{n+1} + \hat{\delta}_{n+1} \frac{4T_{n+1}(y)}{T_{n+2}(y)(b-a)} + q_{n+1}\hat{\delta}_n$, where $y = (a+b)/(b-a)$, $q_{n+1} = (T_n(y))/(T_{n+2}(y))$ and $\hat{\delta}_n$ is given in (4.2). Therefore

$$\hat{r}_{n+2} = q - A\hat{x}_{n+2} = \hat{r}_{n+1} - A\hat{\delta}_{n+1} \frac{4T_{n+1}(y)}{T_{n+2}(y)(b-a)} - Aq_{n+1}\hat{\delta}_n.$$

Replacing $q_{n+1} A\delta_n$ by $q_{n+1} (A\hat{x}_{n+1} + q - q - A\hat{x}_n)$ and collecting terms yields

$$\hat{r}_{n+2} - \hat{r}_{n+1} = -A\delta_{n+1} \frac{4T_{n+1}(y)}{T_{n+2}(y)(b-a)} - q_{n+1}(\hat{r}_n - \hat{r}_{n+1})$$

and

$$A\delta_{n+1} = \frac{T_{n+2}(y)(b-a)}{4T_{n+1}(y)} \left[(\hat{r}_{n+1} - \hat{r}_{n+2}) - q_{n+1}(\hat{r}_n - \hat{r}_{n+1}) \right],$$

which completes the proof.

We shall now develop a method of using iteration (4.2) to solve $Ax=q$ and automatically improve the interval (a,b) . The method will use iteration (4.2) and a sequence of intervals (a_i, b_i) that converge to the optimal interval $(\lambda_{\min}, \lambda_{\max})$. The method is developed from Theorem 4.5 and Lemma 4.1. According to Theorem 4.5, as n increases δ_n^i approaches an eigenvector of $(A+B)^{-1}A$ corresponding to an extreme eigenvalue and μ_i , defined in (4.21), is an approximation to the corresponding eigenvalue. The intervals are improved by replacing a_i with μ_i if it is an approximation to the smallest eigenvalue and replacing b_i with μ_i if it is an approximation to the largest eigenvalue.

Whether μ_i approximates $\lambda_{\min}$ or approximates $\lambda_{\max}$ is determined as follows.

Recall that δ_n^i approaches an eigenvector corresponding to $\lambda_{\min}$ only if $a_i > \lambda_{\min}$ and approaches an eigenvector corresponding to $\lambda_{\max}$ only if $b_i < \lambda_{\max}$. Thus μ_i can approximate $\lambda_{\min}$ only if $a_i > \lambda_{\min}$ and μ_i can approximate $\lambda_{\max}$ only if $b_i < \lambda_{\max}$. Furthermore we are developing a method which will generate a sequence of intervals (a_i, b_i) that converge to $(\lambda_{\min}, \lambda_{\max})$. Therefore we must be able to improve the left endpoint of some intervals and the right endpoint of others. This implies that the condition $a_i > \lambda_{\min}$ and

$b_i < \lambda_{\text{MAX}}$ must be satisfied for each i . If this condition holds it can be concluded that μ_i approximates λ_{min} if $\mu_i \leq a_i$ and approximates λ_{MAX} if $\mu_i \geq b_i$.

Note that the condition $a_i > \lambda_{\text{min}}$ and $b_i < \lambda_{\text{MAX}}$ can be replaced by $a_i \geq \lambda_{\text{min}}$ and $b_i \leq \lambda_{\text{MAX}}$, i.e., $(a_i, b_i) \subset (\lambda_{\text{min}}, \lambda_{\text{MAX}})$. With this relaxed condition, the endpoints of each (a_i, b_i) which are not already optimal can be improved.

The next lemma shows that if $(a_i, b_i) \subset (\lambda_{\text{min}}, \lambda_{\text{MAX}})$ and (a_{i+1}, b_{i+1}) is defined by replacing either a_i or b_i with μ_i , then $(a_{i+1}, b_{i+1}) \subset (\lambda_{\text{min}}, \lambda_{\text{MAX}})$. Therefore if $(a_1, b_1) \subset (\lambda_{\text{min}}, \lambda_{\text{MAX}})$ we may assume $(a_i, b_i) \subset (\lambda_{\text{min}}, \lambda_{\text{MAX}})$ for each i . How to select $(a_1, b_1) \subset (\lambda_{\text{min}}, \lambda_{\text{MAX}})$ is discussed in section 4.5.

Lemma 4.2: If $\mu = \frac{\langle Ay, y \rangle}{\langle (A+B)y, y \rangle}$ where A and $A+B$ are positive definite, then

$$\mu \in [\lambda_{\text{min}}((A+B)^{-1}A), \lambda_{\text{MAX}}((A+B)^{-1}A)] .$$

Proof: This is a restatement of (3.6)

Throughout the remainder of this chapter $P_n(X)$ will denote the polynomial

$$T_n\left(\frac{a+b-2X}{b-a}\right) / T_n\left(\frac{a+b}{b-a}\right), \quad a < b .$$

The method we have developed so far may be summarized as follows:

1. Given a vector $\hat{x}_0^i$ and an interval (a_i, b_i) compute $\hat{x}_n^i$ $n=1, \dots, N$ and $\hat{\sigma}_n^i$ $n=1, \dots, N-1$ by (4.2).
2. Calculate

$$\mu_i = \frac{\langle \hat{\sigma}_{N-1}^i, A \hat{\sigma}_{N-1}^i \rangle}{\langle \hat{\sigma}_{N-1}^i, (A+B) \hat{\sigma}_{N-1}^i \rangle} .$$

3. If $\mu_i \leq a_i$ define $a_{i+1} = \mu_i$ and $b_{i+1} = b_i$
 If $\mu_i \geq b_i$ define $a_{i+1} = a_i$ and $b_{i+1} = \mu_i$
 In either case let $\hat{x}_0^{i+1} = \hat{x}_N^i$ and return to step 1.

We now consider the case $a_i < \mu_i < b_i$. If the approximate eigenvalue is repeatedly in the interval (a_i, b_i) then one of the following conditions holds: Either

$$1. (a_i, b_i) \supset (\lambda_{\min}, \lambda_{\max})$$

or

$$2. P_N(\lambda_{\min}) \cong |P_N(\lambda_{\max})|$$

Remark: We can choose $(a_1, b_1) \subset (\lambda_{\min}, \lambda_{\max})$ and thus assure $(a_i, b_i) \subset (\lambda_{\min}, \lambda_{\max})$. However we will present a heuristic method of selecting the initial interval that does not guarantee $(a_1, b_1) \subset (\lambda_{\min}, \lambda_{\max})$ (see Section 4.5). For this reason it is necessary to consider case 1 above. In either case the vectors $\hat{x}_n$ will converge to the solution of $Ax=q$. In the first case the new method reduces to the method of Chapter 3. The initial interval can be chosen so that if this case occurs, convergence is rapid (see section 4.5). In the second case convergence holds since $0 < P_n(\lambda_{\min}) < 1$ for all n (see Fig. 4.1). Nevertheless, there is a complication.

We shall consider two subcases.

$$2a. a_i \cong \lambda_{\min} \quad \text{and} \quad b_i \cong \lambda_{\max}.$$

$$2b. a_i \not\cong \lambda_{\min} \quad \text{and} \quad b_i \not\cong \lambda_{\max}.$$

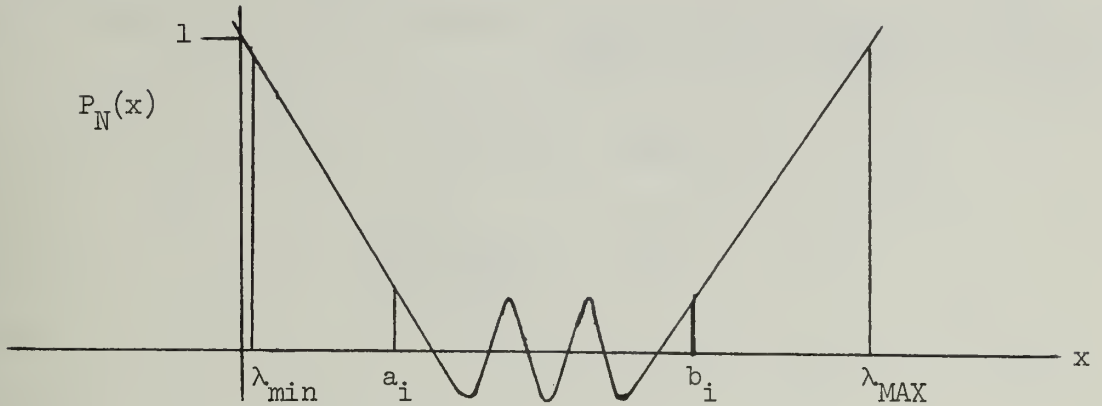
Consider 2a. If $a_i = \lambda_{\min}$ and $b_i = \lambda_{\max}$ then $P_N(\lambda_{\min}) = |P_N(\lambda_{\max})|$.

Thus if the interval is improved so that $(a_i, b_i) \cong (\lambda_{\min}, \lambda_{\max})$, then

$P_N(\lambda_{\min}) \cong |P_N(\lambda_{\max})|$. In this case the interval is nearly optimal and

should be used in the subsequent iterations to compute $\hat{x}_{N+1}^i, \hat{x}_{N+2}^i, \dots$.

In case 2b the situation may be as pictured below.



Convergence of $\{\hat{x}_N^i\}$ will be slow since the interval (a_i, b_i) is a crude approximation to $(\lambda_{\min}, \lambda_{\max})$. Furthermore the vectors $\hat{\delta}_n^i$ do not approach an eigenvector of $(A+B)^{-1}A$. Instead they approach a linear combination of the eigenvectors corresponding to $\lambda_{\min}$ and $\lambda_{\max}$. Note that merely increasing the degree of P , i.e., performing more iterations, will not guarantee convergence of $\{\hat{\delta}_n^i\}$ to an eigenvector of $(A+B)^{-1}A$. It follows that the interval cannot be improved using the method described above.

We shall now describe a technique which will force convergence of $\{\hat{\delta}_n^i\}$ when $P_N(\lambda_{\min}) \cong |P_N(\lambda_{\max})|$. Performing one iteration of (4.1) with $\tau = 1/b_i$ will diminish the components in the error vector and the approximate eigenvector corresponding to the larger eigenvalues. Therefore it will increase the relative size of the components corresponding to the smallest eigenvalues.

This idea can be stated more rigorously as follows. If $N+1$ iterations have been done using the interval (a_i, b_i) and $\hat{\delta}_{N+1}^i = P_N((A+B)^{-1}A)\hat{\delta}_1^i$ where $P_N(\lambda_{\min}) \cong |P_N(\lambda_{\max})|$, then one iteration with $\tau = 1/b_i$ yields

$$\hat{\delta}_{N+2}^i = P_{N+1}^* ((A+B)^{-1}A)\hat{\delta}_1^i$$

where $P_{N+1}^*(X) = (1 - \frac{X}{b_i}) P_N(X)$. Furthermore $P_{N+1}^*(\lambda_{\min}) \neq |P_{N+1}^*(\lambda_{\max})|$.

We could next compute $\hat{r}_{N+1} = q - A\hat{x}_{N+1}$,

$$\hat{\delta}_{N+2} = (A+B)^{-1} \hat{r}_{N+1},$$

$$\mu_{N+2}^i = \frac{\langle \hat{\delta}_{N+2}, A\hat{\delta}_{N+2} \rangle}{\langle \hat{\delta}_{N+2}, \hat{\delta}_{N+1} \rangle}$$

and improve the interval if $\mu_{N+2}^i \notin (a_i, b_i)$.

However, if $\mu_{N+2}^i \in (a_i, b_i)$ we assume that case 2a has occurred and thus want to continue iterating using (a_i, b_i) . There are two methods of using the interval for the next N steps. The first is to define $\hat{x}_0^{i+1} = \hat{x}_{N+1}^i$ and compute $\hat{x}_n^{i+1}$ $n=1, \dots, N$ using $(a_{i+1}, b_{i+1}) = (a_i, b_i)$. The second is to continue using (a_i, b_i) and thus define $\hat{x}_{N+2}^i, \hat{x}_{N+3}^i, \dots, \hat{x}_{2N+1}^i$.

We take the second alternative. Furthermore the calculation of μ_{N+2}^i and the corresponding attempt to improve the interval is omitted. Instead $\hat{x}_{N+1}^i$ is defined using $\tau=1/b_i$ and $\hat{x}_{N+2}^i, \dots, \hat{x}_{2N+2}^i$ are immediately computed. Then

$$\mu_{2N+1}^i = \frac{\langle \hat{\delta}_{2N+1}^i, A\hat{\delta}_{2N+1}^i \rangle}{\langle \hat{\delta}_{2N+1}^i, (A+B)\hat{\delta}_{2N+1}^i \rangle}$$

is computed and used to improve the interval if $\mu_{2N+1}^i \in (a_i, b_i)$. This procedure is chosen for the following reasons. An approximate eigenvalue μ is computed only once every N iterations instead of twice. Next, we know that case 2a will eventually occur in every application of this method and this case is not benefitted by the extra calculation of μ_{N+2}^i . Finally case 2b happens only infrequently. (This can be established by fixing $\lambda_{\min}$ and considering the range of values of $\lambda_{\max}$ that cause case 2b to occur.)

A technical problem now remains. The vectors $\hat{x}_{N+2}^i, \dots, \hat{x}_{N+2}^i$ cannot be computed by simply restarting the iteration using (4.2) since $\delta \hat{x}_N^i$ is not defined. The following will elucidate the problem. Computing $\hat{x}_{n+2}^i$ requires $\hat{x}_{n+1}^i$ and $\delta \hat{x}_n^i$. First, $\delta \hat{x}_{n+1}^i$ is computed and then $\hat{x}_{n+2}^i = \hat{x}_{n+1}^i + \delta \hat{x}_{n+1}^i$. If iteration (4.2) is used to calculate $\hat{x}_1^i, \dots, \hat{x}_N^i$ then the last vector $\delta \hat{x}^i$ to be calculated is $\delta \hat{x}_{N-1}^i$. If $\hat{x}_{N+1}^i$ is computed by using $\tau = 1/b_i$ in iteration (4.1) then it is necessary to also define $\delta \hat{x}_N^i$ before resuming iteration (4.2).

The vector $\delta \hat{x}_N^i$ must be defined so that the iterates $\hat{x}_{N+2}^i, \hat{x}_{N+3}^i, \dots$, converge to x and so that $\delta \hat{x}_{N+2}^i, \delta \hat{x}_{N+3}^i, \dots$, converge to an eigenvector of $(A+B)^{-1}A$. Convergence of $\{\hat{x}_N^i\}$ is established by proving, in the following theorem, that the error vectors $e_n = x - \hat{x}_n^i$ $n \geq N+3$ satisfy $e_n = P_n^*((A+B)^{-1}A)e_0$, where again $P_n^*(X) = (1 - \frac{X}{b_i})P_{n-1}(X)$. Then using the fact that the same polynomial

$$\hat{P}_n((A+B)^{-1}A) = \prod_{j=0}^{n-1} (I - \tau_j(A+B)^{-1}A)$$

appears in

$$e_n = \hat{P}_n((A+B)^{-1}A)e_0$$

and

$$\hat{\delta}_{n+1} = \hat{P}_n((A+B)^{-1}A)\hat{\delta}_1$$

(compare (3.8) and (4.20)) we conclude

$$\hat{\delta}_{n+1} = P_n^*((A+B)^{-1}A)\hat{\delta}_1,$$

and thus $\hat{\delta}$ approaches an eigenvector of $(A+B)^{-1}A$.

Theorem 4.8: Let $\hat{x}_0$ be any initial vector and let $\hat{x}_n$ $n=1, \dots, N$ and

$\delta \hat{x}_n$ $n=0, \dots, N-1$ be defined by (4.2). Let $\hat{x}_{N+1} = \hat{x}_N + \delta \hat{x}_N$,

where

$$\hat{\delta}_{N+1} = \tau(A+B)^{-1}(q-A\hat{x}_N)$$

and let

$$\hat{\delta x}_N = (I - \tau(A+B)^{-1}A)\hat{\delta x}_{N-1}.$$

If $\hat{x}_n$, $n \geq N+2$ is generated by

$$\hat{x}_{n+1} = \hat{x}_n + \hat{\delta x}_n,$$

$$\hat{\delta x}_n = \frac{4T_{n-1}(y)}{(b-a)T_n(y)} (A+B)^{-1}(q-A\hat{x}_n) + \frac{T_{n-2}(y)}{T_n(y)} \hat{\delta x}_{n-1}, \quad (4.22)$$

where $y = (a+b)/(b-a)$. Then

$$e_n = (I - \tau(A+B)^{-1}A)P_n((A+B)^{-1}A)e_0, \text{ where } P_n(X) = \frac{T_n\left(\frac{a+b-2X}{b-a}\right)}{T_n(y)}$$

for $n \geq N+2$.

Proof: The proof will be by induction on e_{n+j} , $j=1, 2, \dots$. The error vectors $e_n = x - x_n$, $n=1, \dots, N$ satisfy $e_n = P_n((A+B)^{-1}A)e_0$ where

$$P_n(X) = \frac{T_n\left(\frac{a+b-2X}{b-a}\right)}{T_n\left(\frac{a+b}{b-a}\right)}.$$

Since $T_{n+1}(X) = 2XT_n(X) - T_{n-1}(X)$ for all n , and $y = (a+b)/(b-a)$ it follows that

$$P_{n+1}(X) = -\frac{T_n(y)4X}{T_{n+1}(y)(b-a)} P_n(X) + P_n(X) + \frac{T_{n-1}(y)}{T_{n+1}(y)} (P_n(X) - P_{n-1}(X)). \quad (4.23)$$

Consequently the polynomial $P_n^*(X) = (1 - \tau X)P_{n-1}(X)$ satisfies

$$P_{n+2}^*(X) = -\frac{T_n(y)4X}{T_{n+1}(y)(b-a)} P_{n+1}^*(X) + P_{n+1}^*(X) + \frac{T_{n-1}(y)}{T_{n+1}(y)} (P_{n+1}^*(X) - P_n^*(X)). \quad (4.24)$$

Let $\hat{x}_{N+1} = \hat{x}_N + \delta_{N+1}$, where $\delta_{N+1} = \tau(A+B)^{-1}(q-A\hat{x}_N)$ and let

$$\delta \hat{x}_N = (I - \tau(A+B)^{-1}A)\delta \hat{x}_{N-1}.$$

$$\text{Obviously } e_{N+1} = P_{N+1}^*((A+B)^{-1}A)e_0. \quad (4.25)$$

Now suppose $\hat{x}_n$ $n=N+2, \dots$, is defined by (4.22). We have

$$\begin{aligned} e_{N+2} &= x - \hat{x}_{N+2} = x - \hat{x}_{N+1} - \delta \hat{x}_{N+1} \\ &= e_{N+1} - \frac{{}_4T_N(y)}{(b-a)T_{N+1}(y)} (A+B)^{-1}(q-A\hat{x}_{N+1}) - \frac{T_{N-1}(y)}{T_{N+1}(y)} \delta \hat{x}_N. \end{aligned} \quad (4.26)$$

But $(q-A\hat{x}_{N+1}) = A(x - \hat{x}_{N+1}) = A(I - \tau(A+B)^{-1}A)e_N$ and

$$\begin{aligned} \delta \hat{x}_N &= (I - \tau(A+B)^{-1}A)(\hat{x}_N - \hat{x}_{N-1}) \\ &= (I - \tau(A+B)^{-1}A)(e_{N-1} - e_N). \end{aligned}$$

Therefore (4.26) can be rewritten as

$$e_{N+2} = (I - \tau(A+B)^{-1}A) \left[e_N - \frac{{}_4T_N(y)}{(b-a)T_{N+1}(y)} (A+B)^{-1}Ae_N + \frac{T_{N-1}(y)}{T_{N+1}(y)} (e_N - e_{N-1}) \right].$$

With (4.23) and $e_N = P_N((A+B)^{-1}A)e_0$ this becomes

$$e_{N+2} = (I - \tau(A+B)^{-1}A)P_{N+1}((A+B)^{-1}A)e_0 = P_{N+2}^*((A+B)^{-1}A)e_0. \quad (4.27)$$

From (4.25) and (4.27) $e_{N+j} = P_{N+j}^*((A+B)^{-1}A)e_0$ for $j=1$ and $j=2$. The

general inductive step follows trivially from (4.24) and $\delta \hat{x}_{N+j} = e_{N+j-1} - e_{N+j}$.

4.5 The Algorithm And Its Convergence

In this section a method of selecting the initial interval (a_i, b_i) will be developed. A precise definition of the algorithm to use (4.2) and

simultaneously improve the endpoints a_1 and b_1 is given. Convergence is also shown. In the next section some empirical results will be given.

Effective application of the methods of the previous section requires the selection of an initial interval $(a_1, b_1) \subset (\lambda_{\min}, \lambda_{\max})$. Lemma 4.3 (below) implies that the only interval guaranteed to satisfy this condition for every problem is the point interval $a_1 = b_1 = 1$. Although the methods of the previous section work when this point interval is used as the initial interval, the resulting iteration may start inefficiently. Indeed, the error vectors e_n may even grow until the interval is improved. Therefore the iterations performed after the interval is improved must annihilate an error vector that is larger than the initial error vector. We shall now consider an alternate method of selection to overcome this inefficiency.

Our goal is to develop a method that requires no estimate of $\lambda_{\min}$ or $\lambda_{\max}$. Since $\lambda_{\min}$ and $\lambda_{\max}$ may be arbitrarily close to unity (see Lemma 4.3) it is impossible to develop a general a priori method of selecting an initial interval $(a_1, b_1) \subset (\lambda_{\min}, \lambda_{\max})$ which avoids the inefficiency of a point interval. We can, however, present a heuristic method of selecting a nondegenerate initial interval that yields a "good" rate of convergence. The interval may fail to satisfy $(a_1, b_1) \subset (\lambda_{\min}, \lambda_{\max})$. However, it will in this case yield a rate of convergence defined by the user to be "good".

Our method of selection is based on the following facts. First, $\lambda_{\min} \leq 1 \leq \lambda_{\max}$. And secondly, the L_2 -norm of the error vector is reduced by a factor $\sigma \leq \frac{1}{T_N\left(\frac{a+b}{b-a}\right)}$ after N iterations when $(a, b) \supset (\lambda_{\min}, \lambda_{\max})$ is used as a constant interval in iteration (4.2). We will show that if

- (1) a_1 and b_1 are chosen from small intervals to the left and right of unity respectively.

and

$$(2) \quad \sigma_N(a_1, b_1) = \frac{1}{T_N\left(\frac{a_1+b_1}{b_1-a_1}\right)} < \epsilon$$

where the user decided that a reduction of the error by a factor less than or equal to ϵ is "good",

then the rate of convergence will be either the "good" rate selected by the user or a rate which is optimal or nearly optimal.

There are three cases:

- (i) $(a_1, b_1) \subset (\lambda_{\min}, \lambda_{\max})$,
- (ii) $(a_1, b_1) \supset (\lambda_{\min}, \lambda_{\max})$ and
- (iii) $a_1 < \lambda_{\min}$ and $b_1 < \lambda_{\max}$ or $a_1 > \lambda_{\min}$ and $b_1 > \lambda_{\max}$.

If $(a_1, b_1) \subset (\lambda_{\min}, \lambda_{\max})$ then the intervals are improved by the methods of the previous section, and approach the optimal interval. Thus the rate of convergence approaches its optimal value. It should be noted that since $(a_1, b_1) \subset (\lambda_{\min}, \lambda_{\max})$ we will have

$$\frac{1}{T_N\left(\frac{a_1+b_1}{b_1-a_1}\right)} \leq \frac{1}{T_N\left(\frac{\lambda_{\min}+\lambda_{\max}}{\lambda_{\max}-\lambda_{\min}}\right)}.$$

Thus the optimal rate of convergence for this problem is slower than the "good" rate selected by the programmer. Also note that since $\lambda_{\min} \leq 1 \leq \lambda_{\max}$, selecting a_1 and b_1 from small intervals to the left and right of unity is a good heuristic for satisfying $(a_1, b_1) \subset (\lambda_{\min}, \lambda_{\max})$. Comments on exact numbers to use as a_1 and b_1 are given below.

If $(a_1, b_1) \supset (\lambda_{\min}, \lambda_{\max})$ then the method never changes the interval and the rate of convergence is the "good" rate selected by the programmer.

Finally, if $a_1 > \lambda_{\min}$ and $b_1 > \lambda_{\max}$ or $a_1 < \lambda_{\min}$ and $b_1 < \lambda_{\max}$ then the endpoint in $(\lambda_{\min}, \lambda_{\max})$ is improved to its optimal value. The endpoint not in $(\lambda_{\min}, \lambda_{\max})$ remains unchanged. As in the case in which both a_1 and b_1 are in $(\lambda_{\min}, \lambda_{\max})$ the rate of convergence is improved. This improved rate approaches a "nearly" optimal rate. To make this more precise consider the case $a_1 > \lambda_{\min}$ and $b_1 > \lambda_{\max}$. The left endpoints, a_i , will be decreased to approximately $\lambda_{\min}$. The right endpoints, however, will never be improved. The rate of convergence will approach the rate obtained using the constant interval $(\lambda_{\min}, b_1)$. But b_1 was chosen in a small interval to the right of unity, say $b_1 = 1 + \delta$. In addition $\lambda_{\max} \geq 1$ so that $|b_1 - \lambda_{\max}| = b_1 - \lambda_{\max} \leq \delta$.

Thus, intuitively, although $b_1 > \lambda_{\max}$ b_1 is not a gross overestimate of $\lambda_{\max}$ and convergence is not significantly slower than the optimal rate.

Remarks: The values of a_1 and b_1 are selected to yield a "good" rate of convergence as a precaution for cases in which $(a_1, b_1) \not\subset (\lambda_{\min}, \lambda_{\max})$. The only purpose in selecting this "good" rate is to prevent the selection of too large an interval with an unacceptable rate.

It is impossible to create a rate of convergence that is better than the optimal rate of the problem. That is, selecting (a_1, b_1) such that

$$\sigma_N(a_1, b_1) = \frac{1}{T_N\left(\frac{a_1 + b_1}{b_1 - a_1}\right)}$$

is very small cannot improve the rate of convergence. In fact, as

$\sigma_N(a_1, b_1)$ goes to zero the interval becomes a point, which is the exact condition we are attempting to avoid.

Our experience is that $a_1 \in (1/4, 8/10)$ and $b_1 \in (1.25, 3)$ yield "reasonable" initial intervals. The interval $(1/3, 5/2)$ has proved to be effective for problems in which nothing is known about $\lambda_{\min}$ or $\lambda_{\max}$. In Chapter 5 a value for a_1 in some special cases is derived. In these cases the interval $(a_1, 5/2)$ has been effective.

The following lemma establishes the basis for the heuristic method of selecting the initial interval.

Lemma 4.3: Let A be as in (2.6) and let B be given by (2.7). Then 1 is an eigenvalue of $(A+B)^{-1}A$. Furthermore for every $\epsilon > 0$ there is some A such that $1 - \lambda_{\min}((A+B)^{-1}A) < \epsilon$ and $\lambda_{\max}((A+B)^{-1}A) - 1 < \epsilon$.

Proof: If B is given by (2.7) then the first row and column are null. Hence B has a zero eigenvalue. Let w be the corresponding eigenvector. Then $(A+B)w = Aw$, i.e., $(A+B)^{-1}Aw = w$.

To see that $\lambda_{\min}$ and $\lambda_{\max}$ can be arbitrarily close to unity consider the following. As the elements $B_{j,k}$ in matrix A (2.4) tend to zero, elements $b_{j,k}$ and $f_{j,k}$ in L and U (2.6) also tend to zero. In the limiting case $B_{j,k} \equiv 0$, $b_{j,k} \equiv 0$ and $f_{j,k} \equiv 0$. Moreover $LU = A$, that is, B is the null matrix. In this case $(A+B)^{-1}A = A^{-1}A = I$. Hence every eigenvalue of $(A+B)^{-1}A$ is unity. Since the eigenvalues vary continuously with the elements of the matrix, for a non-null B with arbitrarily small elements $\lambda_{\min}$ and $\lambda_{\max}$ may be arbitrarily close to unity.

Remark: This lemma is valid for the factorizations due to Kim [11] and Dupont, Kendall and Rachford [5].

We shall now give a statement of the algorithm.

The Algorithm.

1. Select an initial interval (a_1, b_1) according to one of the following rules.

(a) Define (a_1, b_1) according to the method presented in the first part of this section. For example $a_1 \in (.25, .75)$ and $b_1 \in (1.5, 3.0)$. In practice, $a_1 = 1/3$, $b_1 = 5/2$ have been effective.

(b) If possible use the approximation of a_1 developed in Chapter 5, Theorem 5.3. Select b_1 as in (a).

(c) Let $a_1 = b_1 = 1$.

Comment: Although (c) yields a convergent and practical iteration, it is recommended that (a) or (b) be used to define the initial interval.

2. Select an initial vector x_0 and some $\epsilon > 0$ which is a bound on the maximum acceptable error, measured by the norm of the residual.
3. Select an integer N which is the number of iterations to perform before attempting to correct the current interval.

Comment: As N is increased the convergence of a_i and b_i to $\lambda_{\min}$ and $\lambda_{\max}$ is improved but the convergence of $\hat{x}_n$ to x is retarded if the current values of a_i and b_i are crude approximations to $\lambda_{\min}$ and $\lambda_{\max}$. In experiments using the algorithm, values of $N = 5, 6, 7, 8, 9$, and 10 were equally effective.

4. Execute N steps of iteration (4.2) generating

$$\hat{x}_N^i, \delta_{N-1}^i \text{ and } \delta \hat{x}_{N-1}^i,$$

where i corresponds to (a_i, b_i) .

5. If $\|\hat{r}_N^i\| = \|q - Ax_N^i\| < \epsilon$ then stop, otherwise compute

$$\mu_i = \frac{\langle A_{N-1}^i, \hat{\delta}_{N-1}^i \rangle}{\langle (A+B) \hat{\delta}_{N-1}^i, \hat{\delta}_{N-1}^i \rangle}.$$

6. i. If $\mu_i < a_i$, define $a_{i+1} = \min \{ \min_{j \leq i} a_j, \mu_i \}$ and $b_{i+1} = b_i$.

Add one to i and go to 4.

ii. If $\mu_i > b_i$, do one iteration of (4.1) with $\tau = 1/\mu_i$. Define $a_{i+1} = b_i$ and $b_{i+1} = \mu_i$. Add one to i and go to 4.

Comment: The definition of the next interval when $\mu_i > b_i$ is different from the definition when $\mu_i < a_i$ for the following reason. Components of the error vector corresponding to eigenvalues $\lambda_k > b_i$ will have been amplified if $\left| P_N^i(\lambda_k) \right| > 1$, where the i indicates correspondence to the interval (a_i, b_i) . When $\mu_i > b_i$, the next interval is defined to emphasize the annihilation of the dominant components of the error. The components which became dominant when $\mu_i < a_i$, on the other hand, can become dominant only if they are not diminished as rapidly as other components. They are never amplified since $\left| P_N(x) \right| < 1$ for all $x \in (0, a_i)$. Therefore when $\mu_i < a_i$ it is unnecessary to concentrate on the reduction of the dominant components.

iii. If $\mu_i \in [a_i, b_i]$, do one iteration with $\tau = 1/b_i$. Add one to i and do N iterations according to (4.22), then go to 5.

Comment: This requires essentially no new code since (4.22) is identical to (4.2) except that the subscripts on the

right side of the equations are decreased. Furthermore the code required to form $(A+B)^{-1}(q-A\hat{x}_N^i)$ and $(I-\tau(A+B)^{-1}A)\delta x_N^i$ before proceeding with (4.22) already exists in the code to perform (4.2).

The purpose of this step is to correct the condition

$$P_N(\lambda_{\min}) \cong \left| P_N(\lambda_{\max}) \right| \text{ as described preceding Theorem 4.8.}$$

Convergence: The convergence of the algorithm is established by considering three types of initial intervals (a_1, b_1) .

1. If $(a_1, b_1) \supset (\lambda_{\min}, \lambda_{\max})$ then the algorithm reduces to the method of Dupont, Kendall and Rachford [5] and is therefore convergent.
2. If $(a_1, b_1) \not\supset (\lambda_{\min}, \lambda_{\max})$ and $P_N(\lambda_{\min}) \cong \left| P_N(\lambda_{\max}) \right|$ then Theorem 4.5 guarantees the convergence of the intervals (a_i, b_i) to an interval (a, b) such that $a \cong \lambda_{\min}$ or $b \cong \lambda_{\max}$. Furthermore if $(a_1, b_1) \subset (\lambda_{\min}, \lambda_{\max})$ then the (a_i, b_i) approach $(\lambda_{\min}, \lambda_{\max})$ and the rate of convergence of $\hat{x}_n$ to x approaches the optimal rate.
3. The case $P_N(\lambda_{\min}) \cong \left| P_N(\lambda_{\max}) \right|$ is trivially disposed of since $0 < P_N(\lambda_{\min}) < 1$.

Work: The following considerations show that if N is the number defined in step 3 of the algorithm the amount of additional computation per iteration required by the new method as compared to the method of Dupont, Kendall and Rachford, (4.2), is approximately $\frac{1}{11N}$ times the number of computations in one iteration of their method.

There are two differences in computation of the methods. The first is that our algorithm computes an approximate eigenvalue on every N^{th} iteration. This requires 2 vector additions and the accumulation of 2

inner products. Therefore the number of extra calculations required per iteration to compute the approximate eigenvalue is $(3K \text{ additions} + 2K \text{ multiplications})/N$, where K is the dimension of A .

The second difference is that after every N^{th} iteration the new algorithm employs a new interval. This creates a difference in the work performed because the first iteration with each interval is a special case. In general

$$\hat{x}_n^i = \frac{4T_n(y)}{(b_i - a_i)T_{n+1}(y)} \hat{x}_n^i + \frac{T_{n-1}(y)}{T_{n+1}(y)} \hat{x}_{n-1}^i ,$$

where $\hat{x}_n^i = (A+B)^{-1} \hat{r}_{n-1}^i$ and $y = (a_i + b_i)/(b_i - a_i)$. When $n=0$, however,

$$\hat{x}_0^i = \frac{2}{b_i + a_i} \hat{x}_0^i , \quad \hat{x}_0^i = (A+B)^{-1} \hat{r}_0^i .$$

is computed. In every case $\hat{x}_{n+1}^i$ is defined as $\hat{x}_n^i + \hat{x}_n^i$. Thus the first iteration with each new interval requires K fewer additions and K fewer multiplications. The additional calculations due to these differences is

$$(2K \text{ additions} + K \text{ multiplications})/N$$

The number of computations in one iteration of (4.1) or (4.2) without the calculation of the approximate eigenvalue is

$$(13K \text{ additions} + 11K \text{ multiplications}).$$

Considering only the multiplications we see that

$$\left[\begin{array}{l} \text{the additional calculations} \\ \text{per iteration of the new} \\ \text{algorithm} \end{array} \right] \cong \frac{1}{11N} \left[\begin{array}{l} \text{calculations per iteration} \\ \text{of (4.1) or (4.2)} \end{array} \right] .$$

4.6 Experimental Results

Four distinct problems will be studied in this section. Stone [16] and Taranto [17] have also made studies using the type of problems

considered here. They compared the amount of work required to solve the system of equations $Ax=q$ using first Stone's factorizations and then using ADI. Taranto also compared Stone's symmetric factorization to Stone's nonsymmetric factorization and studied the effect of changing the parameters α and τ (see Chapter 2 for definitions of α and τ).

The following tests make two comparisons. First the number of iterations and the amount of work required to solve systems of equations using Stone's symmetric factorization and the algorithm developed in this chapter is compared with the number of iterations and the work required when the problems are solved using Stone's factorization with a single best parameter. Next the comparisons are made for problems solved by Stone's symmetric factorization with the new algorithm and both of Stone's factorizations used with Stone's algorithm to vary the auxiliary matrix, [15, 16].

Problems a and c described below are identical to problems studied by Stone and Taranto. Problem b is similar to a problem they studied but differs in the ratio of the functions $a_1(x)$ and $a_2(x)$. (see Chapter 2, where a_1 and a_2 are defined.) This change was made because the problem studied by Stone and Taranto converges so rapidly using the methods under consideration that no significant difference in the relative rates of convergence results. The last problem studied involves random numbers in the matrix A . Since the random numbers generated in the problem below are not the same as those generated in the previous studies, no comparison of results is attempted. In fact the form of the problem is changed slightly since the comparison is impossible. In the problem studied below the functions a_1 and a_2 of (2.1) are random throughout the region. In the problem studied

by Stone and Taranto the functions were constant on subregions but were otherwise random.

Each problem is solved on the unit square, $\{(x_1, x_2) : 0 \leq x_i \leq 1 \ i=1,2\}$, with zero boundary condition, i.e., $\Psi(x)$ of (2.1) is identically zero. The solution x is chosen to be

$$x_{j,k} = \cos(j\pi h) \cdot \cos(k\pi h) \quad j,k=1, \dots, 30,$$

where $h=1/(30+1)$. Similar tests were run with cosine replaced by sine. The results were almost identical and therefore are not presented. The vector q of (2.2) is computed from

$$(Ax)_{j,k} = q_{j,k}.$$

The initial vector, x_0 , is chosen to be the zero vector.

Problem a. This problem is the model problem, defined by letting $a_1(x)$ and $a_2(x)$ be identically -1. The initial estimate of $\lambda_{\min}$, suggested by the analysis of Chapter 5 is $\frac{1}{2}$.

Problem b. For this problem, called the generalized model problem, $a_1(x) = -1/9$ and $a_2(x) = -1$. The initial estimate of $\lambda_{\min}$ as suggested in Chapter 5 is $5/8$.

Problem c. This problem is defined by a heterogeneous region containing homogeneous subregions see Figure 4.2. In region A, the ratio of $a_1(x)$ to $a_2(x)$ is unity with $a_1(x) = -1$. In region B, this ratio is 10 with $a_1(x) = -1$; in C it is 0.1 with $a_1(x) = -1/10$ and in D it is 1 with $a_1(x) = -1/10$. The theory of Chapter 5 does not yield an estimate of $\lambda_{\min}$ in this case.

Problem d. In this problem random values of $a_1(x)$ are chosen from the set $\{z : 0 < z < 1\}$.

Each of the four problems is solved several times with N of step 3 of the algorithm equal to 6 and various numbers selected as initial endpoints a_1 and b_1 . The set of endpoints was chosen so that the performance of the algorithm could be examined for both good and poor values. The endpoints were chosen to create the following situations.

1. a_1 a poor approximation to $\lambda_{\min}$;
2. b_1 a poor approximation to $\lambda_{\max}$;
3. a_1 and b_1 poor approximations to $\lambda_{\min}$ and $\lambda_{\max}$ respectively.
4. a_1 and b_1 chosen according to the suggestions in step 1 parts a and b of the algorithm
5. Values of a_1 between the poor and the suggested values with $b_1 = 2.5$.

The problems were then solved using a single best value of τ . Theoretically this value of τ is $2/(\lambda_{\min} + \lambda_{\max})$; however, this value was determined experimentally. Values of $\lambda_{\min}$ and $\lambda_{\max}$ were not computed directly.

The results of these tests are displayed in the following graphs. Only linear approximations to the curves are shown in order to simplify comparisons.

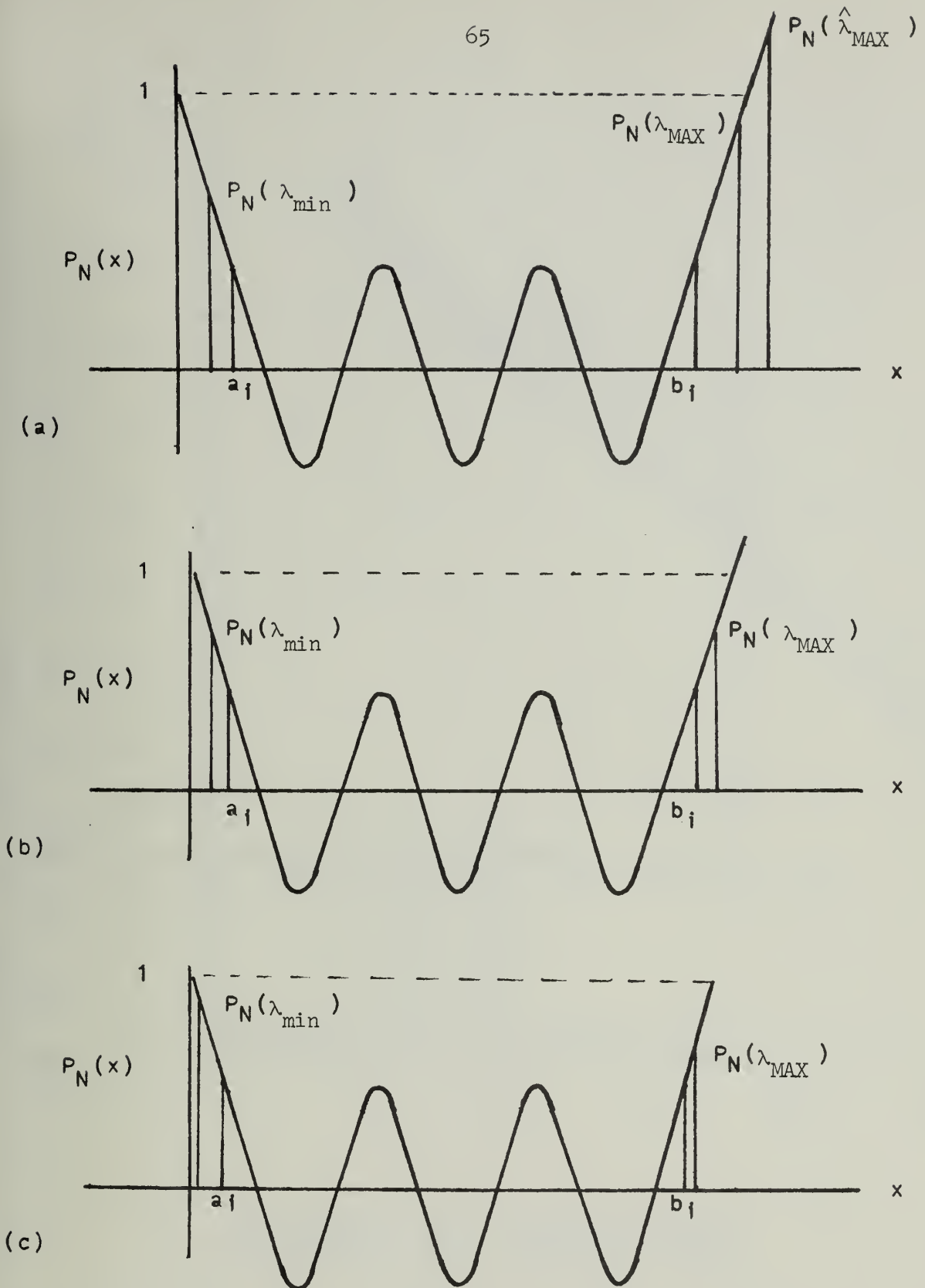


Figure 4.1.

- (a) Two cases: In Both Cases the Component Corresponding to λ_{MAX} Is Dominant. In One Case the Error Is Increasing, in the Other It Is Decreasing.
- (b) $|P_N(\lambda_{min})| = |P_N(\lambda_{MAX})|$ and the Error Is Decreasing.
- (c) The Component Corresponding to λ_{min} Is Dominant and the Error Is Decreasing.

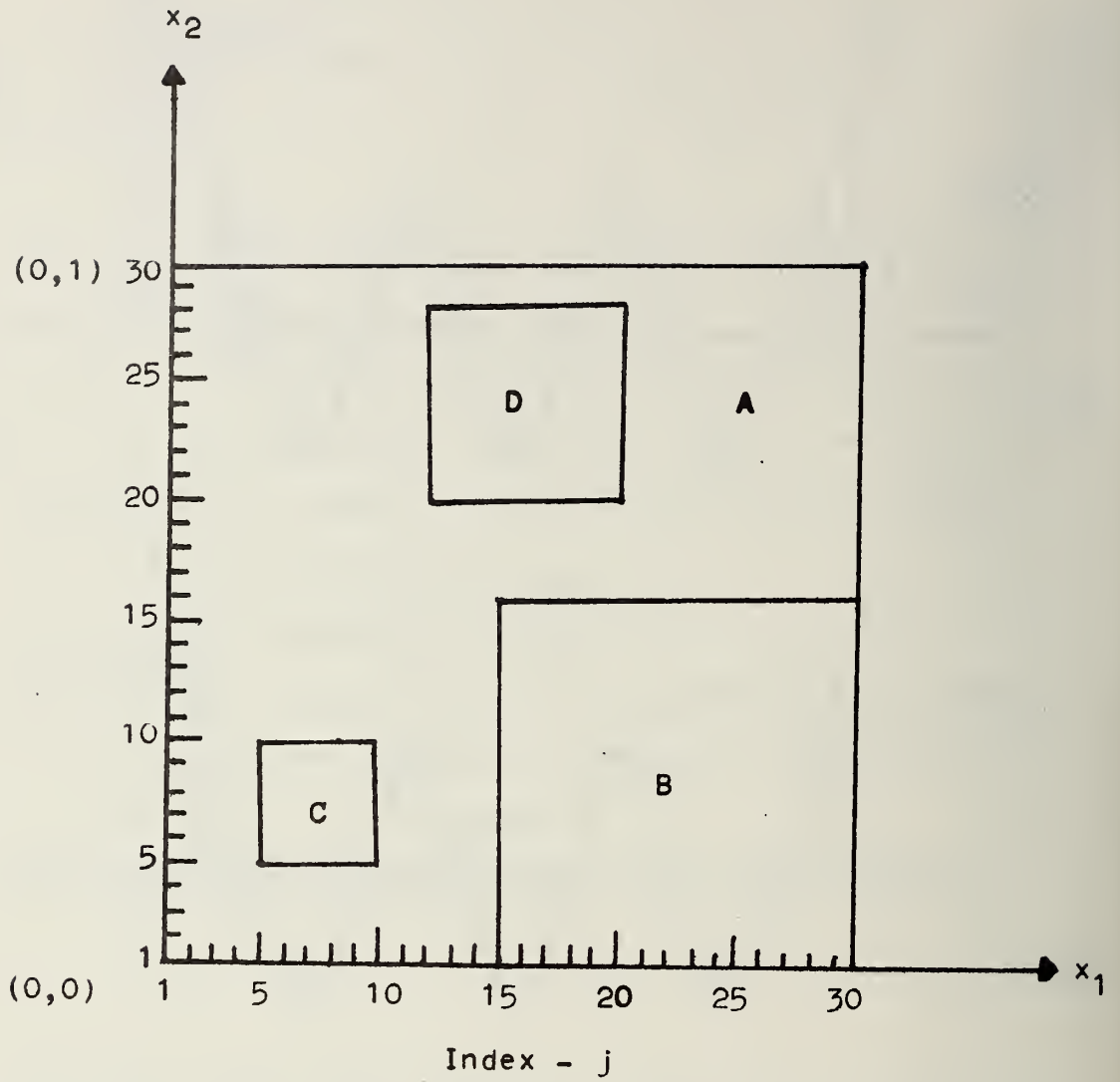


Figure 4.2. Heterogeneous Test Problem Geometry
(Region Defined on Unit Square)

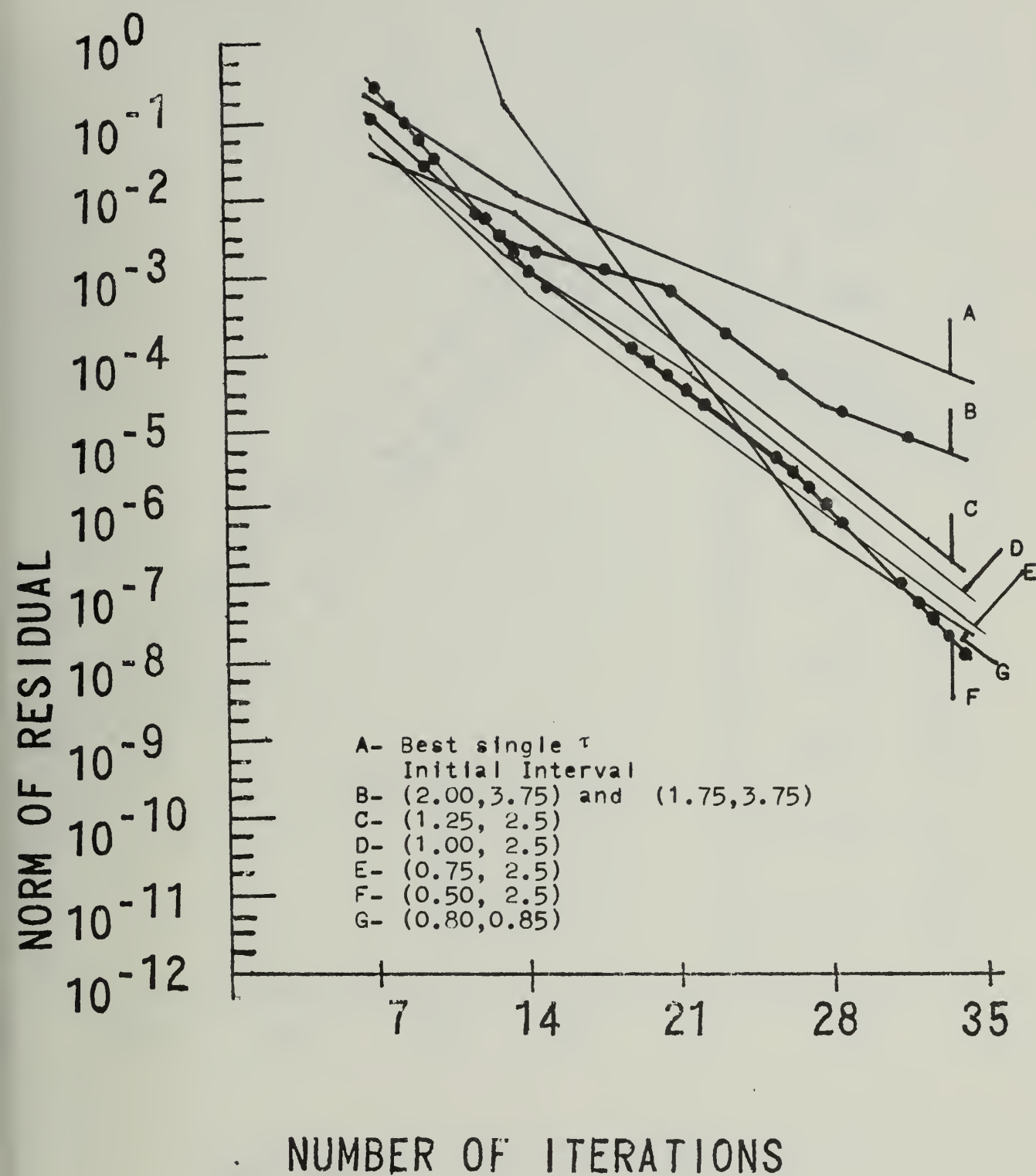


Figure 4.3. Model Problem

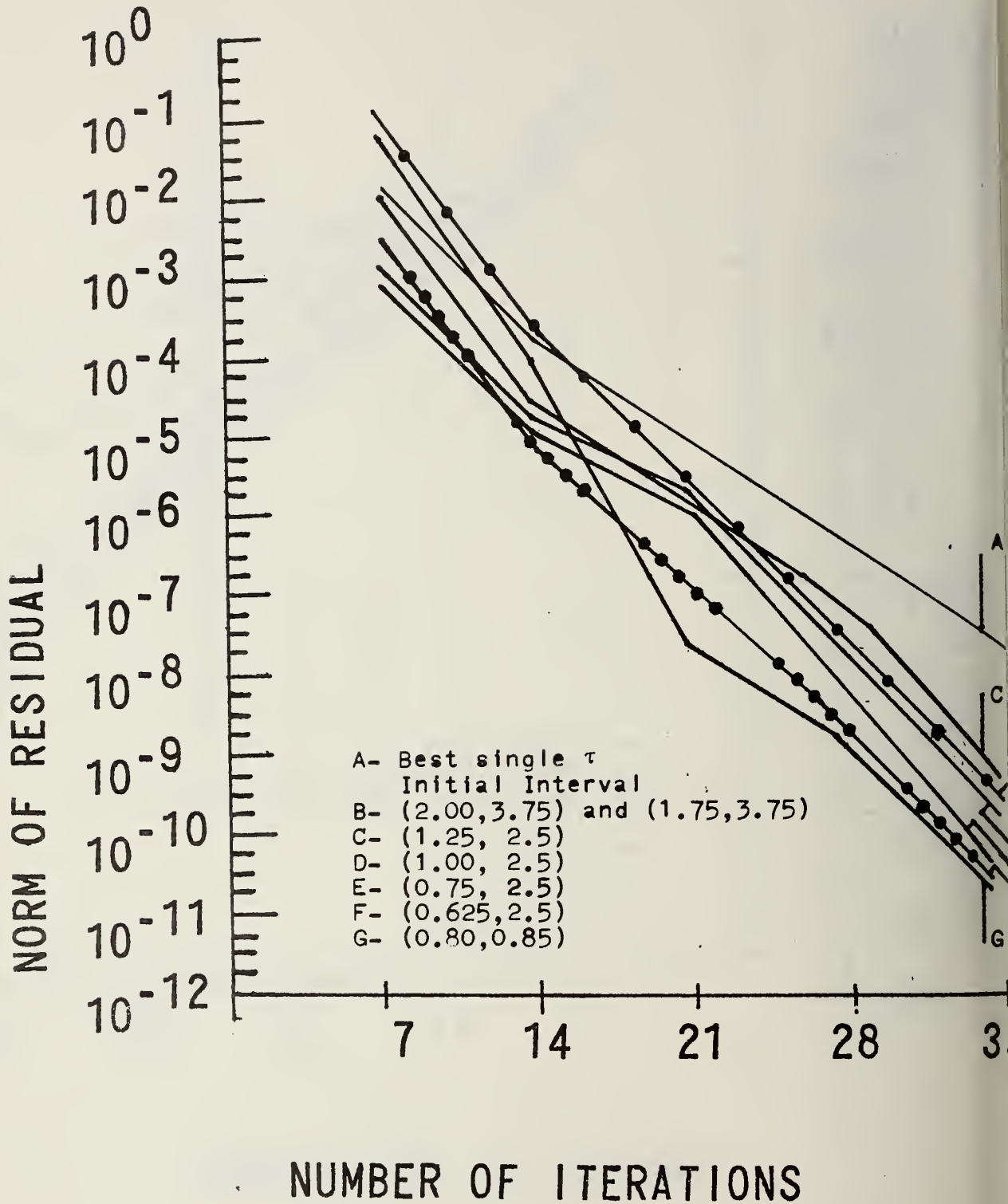


Figure 4.4. Generalized Model Problem

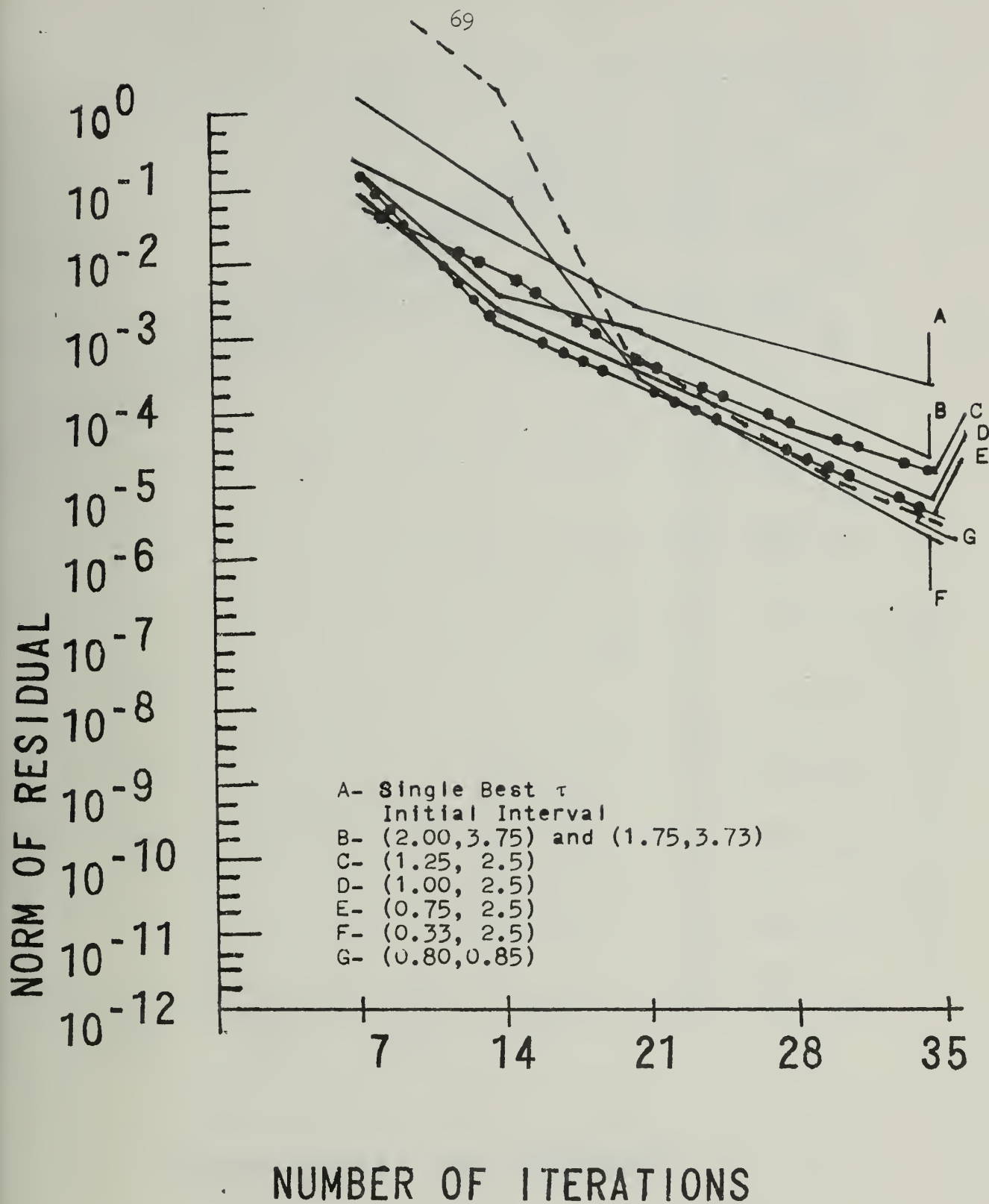


Figure 4.5. Heterogeneous With Homogeneous Subregions

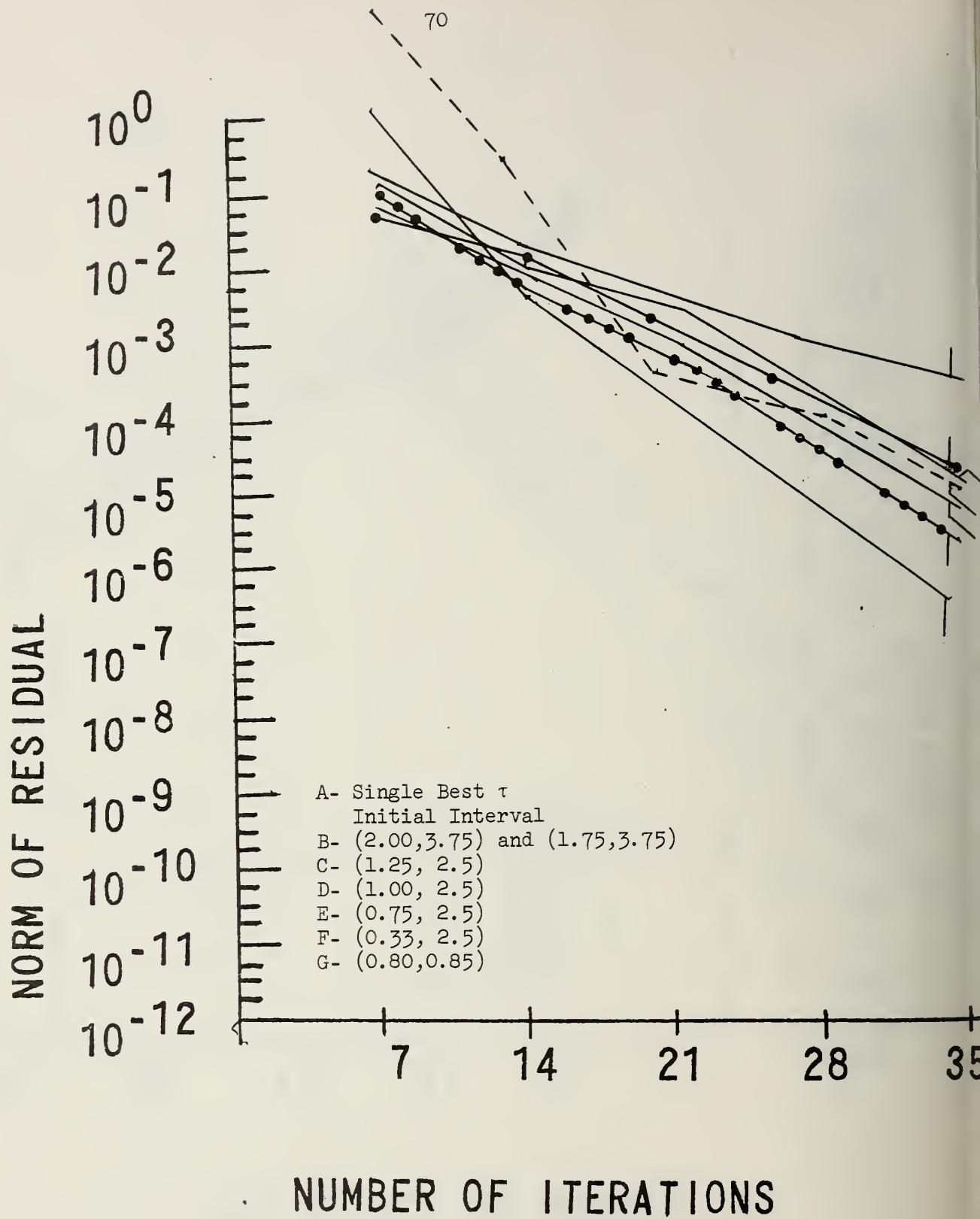


Figure 4.6. Random Numbers in the Matrix A

The next four graphs display the results of problems a and c above when solved by the following three methods.

1. The algorithm of Chapter 4 with the initial interval $(1/3, 5/2)$.
2. Stone's nonsymmetric factorization implemented with a sequence of α 's as suggested in [16], namely

$$\alpha_m = \alpha_{\text{MAX}} = 1 - \min \left[\frac{2h^2}{1 + a_2/a_1}, \frac{2h^2}{1 + a_1/a_2} \right] \quad (4.31)$$

and $\alpha_1 = \alpha_{\text{min}} = 0$; the intermediate parameters

α_i , $i=2, \dots, m-1$ are computed from

$$\alpha_i = 1 - (1 - \alpha_{\text{MAX}})^{\frac{i-1}{m-1}}.$$

a_1 and a_2 in (4.31) are the constant values of $a_1(x)$ and $a_2(x)$ respectively (see Chapter 2 for definition of $a_1(x)$ and $a_2(x)$). In problem c there is a value of α_i defined for each subregion using the local constant values of $a_1(x)$ and $a_2(x)$. For the tests presented here $m=9$. The α_i are used cyclicly with each α_i used on two consecutive iterations. On the odd numbered iterations the gridpoints are ordered in the normal ordering; i.e., the gridpoint (i,j) (see Chapter 2) precedes the gridpoint (k,j) if $i < k$ and precedes the gridpoint (l,m) if $j < m$. On the even numbered iterations the reverse ordering is used; i.e.,

the gridpoint (i,j) precedes the gridpoint (k,j)
 if $i < k$ and precedes the gridpoint (l,m) if $j > m$.

3. Stone's symmetric factorization implemented with a sequence of α 's as in 2.

The results for methods 2 and 3 are taken from the study done by Taranto [17].

The graphs in Figures 4.7 and 4.8 have the reduction of the error is plotted against the number of iterations. The reduction of the error is calculated as

$$\|e_0\| / \|e_n\|$$

where e_0 is the error in the initial vector and e_n is the error in the n^{th} iterate. The graphs in Figures 4.9 and 4.10 have the reduction of the error plotted against the work performed. One unit of work is taken to be the amount required to do one iteration of (4.2). It is assumed that all the auxiliary matrices, B^α , are calculated only once and that the amount of work needed to calculate the elements of the factor L^α and U^α is equivalent to the amount required to do one iteration.

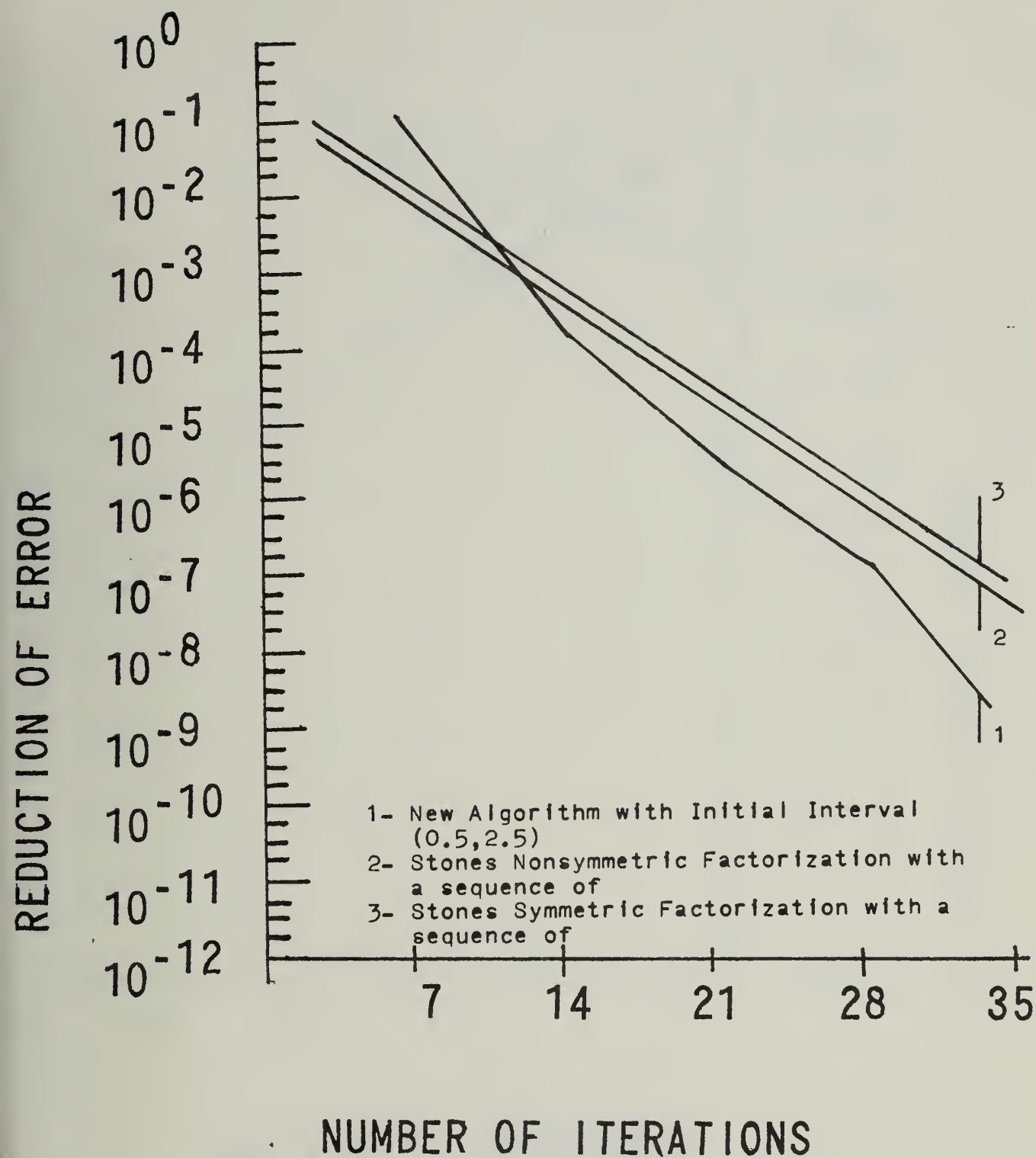
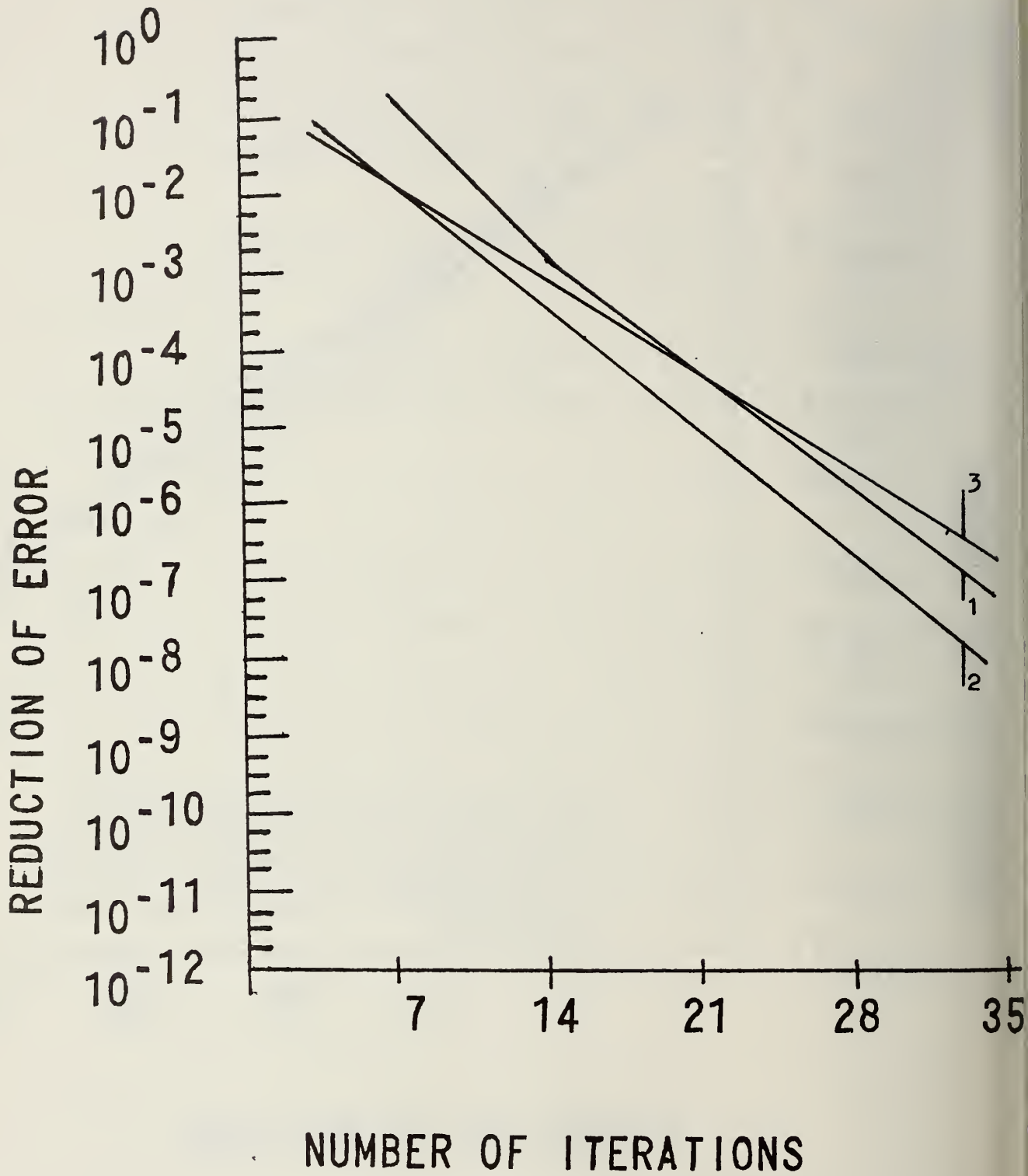


Figure 4.7. Model Problem



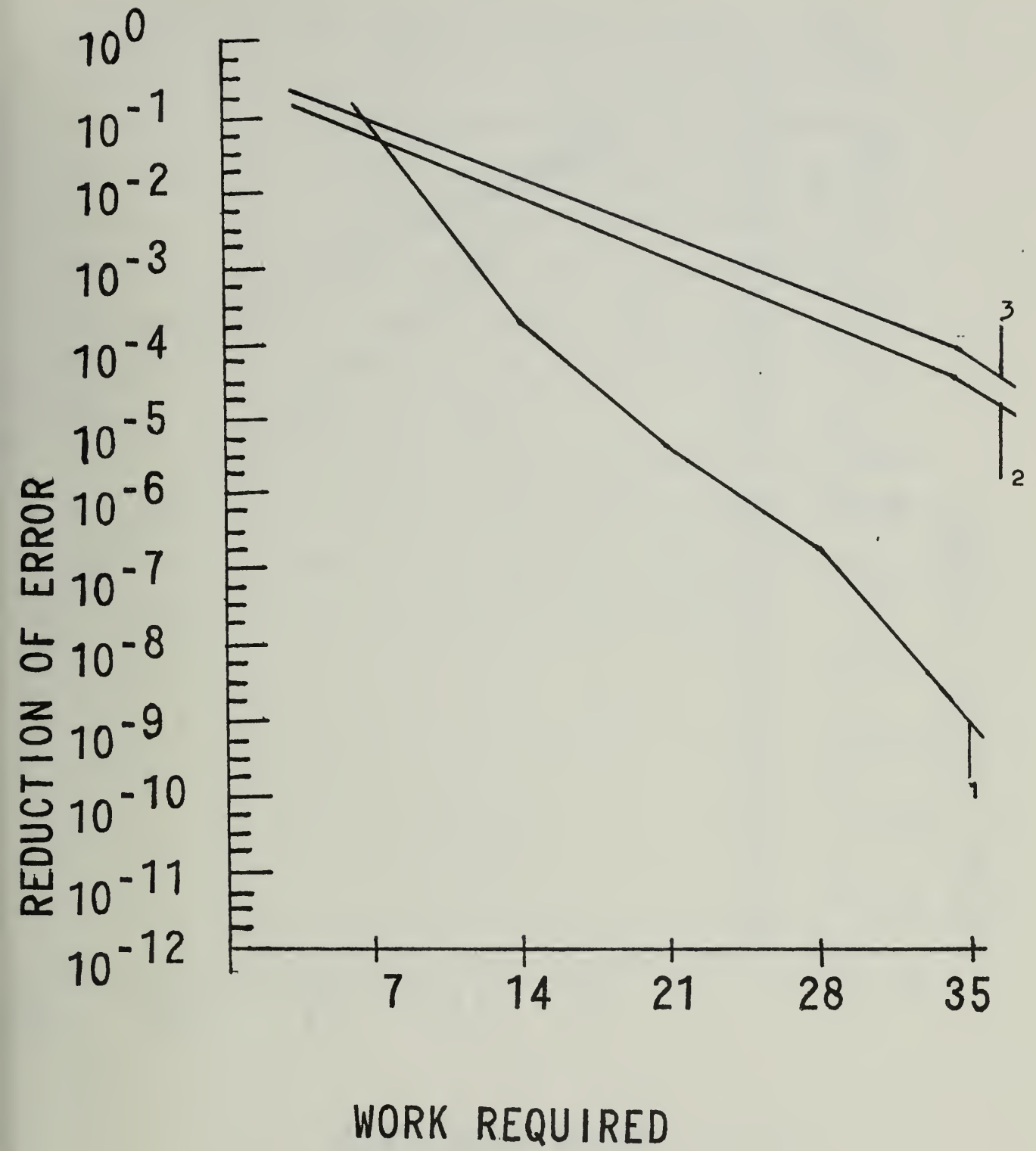


Figure 4.9. Model Problem

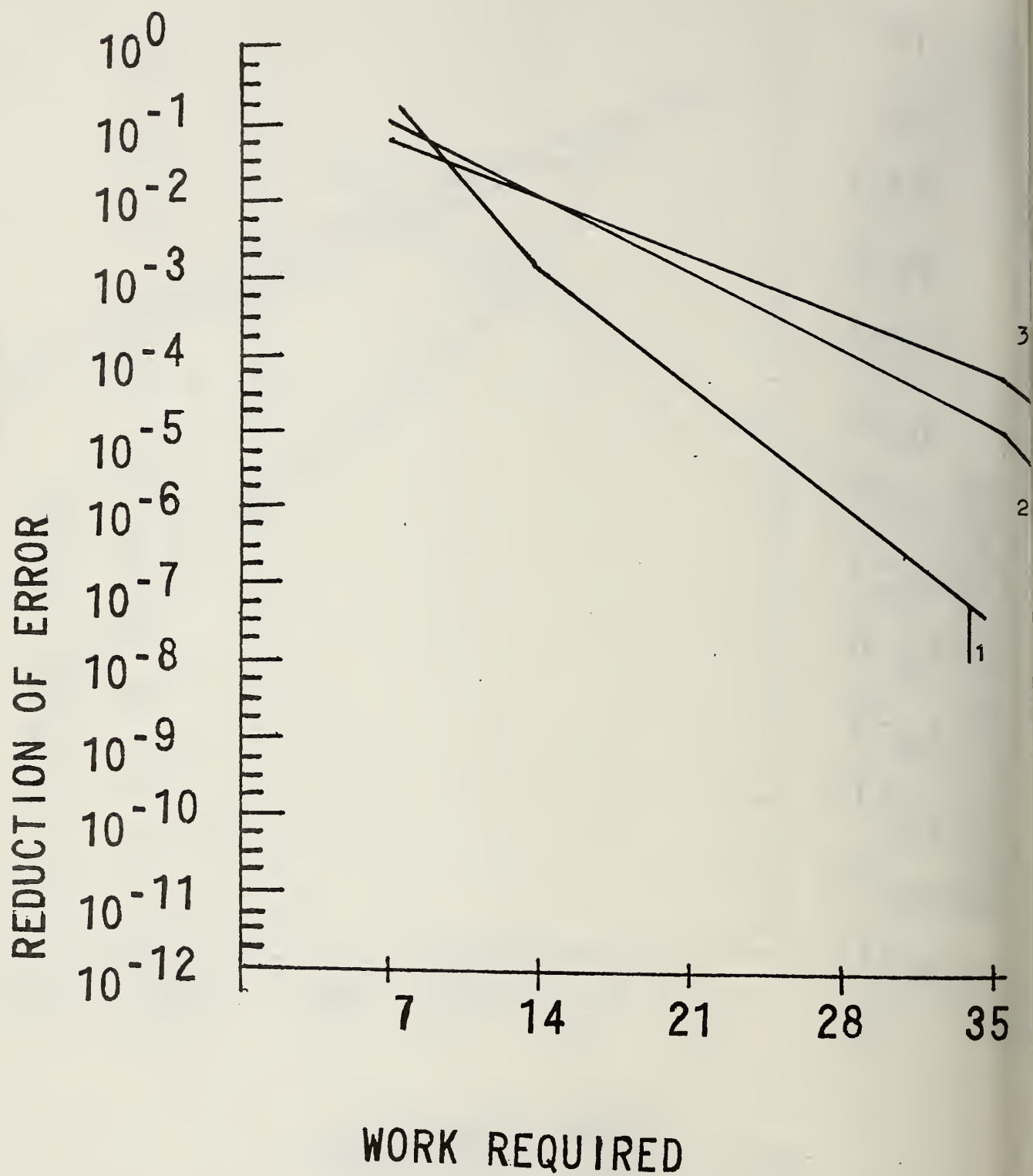


Figure 4.10. Heterogeneous with Homogeneous Subregions

Conclusions: The tests comparing the algorithm developed in this chapter to the use of a single best τ clearly indicate that fewer iterations are required when the new algorithm is used. Since the tests were performed with N of step 3 in the algorithm equal to 6, the work estimate of the last section implies that the additional work required by the new algorithm is approximately $1/66$ of the work performed by the iteration which doesn't change the interval. Thus the same graphs may be used to conclude that the new algorithm requires less work than the iteration using a single best parameter.

The graphs of Figures 4.7 and 4.8 indicate that in some cases the required number of iterations of the algorithm developed is greater than the number required by Stone's algorithm. However, Figures 4.9 and 4.10 indicate less work is required by the new algorithm.

5. AN ESTIMATE OF $\lambda_{\min}$

Throughout this chapter the following notation will be used. B will denote the auxiliary matrix defined using (2.6) and (2.7) with $\alpha = 1$. The quantities $\lambda_{\min} = \lambda_{\min}((A+B)^{-1}A)$, $b_{j,k}$, $c_{j,k}$, $d_{j,k}$, $e_{j,k}$, $f_{j,k}$, $C_{j,k}$ and $G_{j,k}$ will be as in (2.7). $B_{j,k}$, $D_{j,k}$ and $E_{j,k}$ will be as in (2.4). When $B_{j,k}$, $D_{j,k}$ and $E_{j,k}$ have constant values, except $B_{j,1} = 0$ and $D_{1,k} = 0$, they will be denoted by B , D and E respectively. Whether B is the auxiliary matrix of (2.6) or the scalar constant $B_{j,k}$ of (2.4) will be obvious from the use. Finally, the elements of the matrix B will be referred to as $\hat{b}_{i,j}$ to avoid any confusion with $b_{j,k}$ of (2.6).

5.1 The Limiting Values Of $\hat{b}_{i,j}$

In this section it will be shown that if $B_{j,k} \equiv B$ and $D_{j,k} \equiv D$, except for $D_{1,k} = 0$ and $B_{j,1} = 0$, then the elements $\hat{b}_{i,j}$, of B , approach a limit as both subscripts increase. In particular $G_{j,k} \rightarrow G = \sqrt{BD}$ as j and k increase. This will be used in the next section to derive an estimate of $\lambda_{\min}$.

The following lemma will establish some useful properties of $A+B$. Although we are only concerned with the case $\alpha = 1$, the lemma is valid for all $\alpha \in [0,1]$.

Lemma 5.1: The quantities defined in (2.6), (2.7) exist and satisfy the following relations:

$$(a) \quad b_{j,k} \leq B_{j,k} \leq 0$$

$$(b) \quad c_{j,k} \leq D_{j,k} \leq 0$$

$$(c) \quad d_{j,k} < 0$$

$$(d) \quad -1 \leq e_{j,k} \leq 0$$

$$(e) \quad -1 \leq f_{j,k} \leq 0$$

$$(f) \quad e_{j,k} + f_{j,k} + 1 \geq 0$$

$$(g) \quad b_{j,k} = d_{j,k-1} f_{j,k-1}$$

$$(h) \quad c_{j,k} = d_{j-1,k} e_{j-1,k}$$

$$(i) \quad d_{j,k} + B_{j,k} + D_{j,k} - b_{j,k} e_{j,k-1} - c_{j,k} f_{j-1,k} \geq 0$$

Proof: Parts (a) through (h) were proved by Bracha [1]. To prove (i) we have,

$$\begin{aligned} & d_{j,k} + B_{j,k} + D_{j,k} - b_{j,k} e_{j,k-1} - c_{j,k} f_{j-1,k} \\ &= -(B_{j,k} + D_{j,k}) - b_{j,k} e_{j,k-1} - c_{j,k} f_{j-1,k} + B_{j,k} - b_{j,k} \\ &+ D_{j,k} - c_{j,k} - b_{j,k} f_{j,k-1} - c_{j,k} e_{j-1,k} \\ &= -b_{j,k}(1 + e_{j,k-1} + f_{j,k-1}) - c_{j,k}(1 + e_{j-1,k} + f_{j-1,k}). \end{aligned}$$

Using (a), (b) and (f) yields

$$d_{j,k} + B_{j,k} + D_{j,k} - b_{j,k} e_{j,k-1} - c_{j,k} f_{j-1,k} \geq 0$$

which completes the proof.

Before demonstrating that $G_{j,k} = c_{j,k} f_{j-1,k}$ tends to $\sqrt{BD}$ as j and k increase, it is shown that for fixed k the quantities of (2.7) approach limits as j increases.

Lemma 5.2: Let $B_{j,k} \equiv B$ and $D_{j,k} \equiv D$, except $B_{j,1} = 0$ and $D_{1,k} = 0$.

For fixed k the elements $b_{j,k}$, $c_{j,k}$, $d_{j,k}$, $e_{j,k}$, $f_{j,k}$, $C_{j,k} = b_{j,k} e_{j,k}$ and $G_{j,k} = c_{j,k} f_{j-1,k}$ tend to limits as j increases. These limits will be called b_k , c_k , d_k , e_k , f_k , C_k and G_k respectively.

Proof: The proof will be done by an induction on k . For $k = 1$ we have the following.

$$b_{j,1} \equiv B_{j,1} = 0 \text{ and } c_{j,1} \equiv D_{j,1} = D.$$

$$d_{j,1} = -2(B+D) - D e_{j-1,1} \text{ and } e_{j,1} = D/d_{j,1}$$

hence

$$d_{j,1} = -2(B+D) - D^2/d_{j-1,1} \tag{5.1}$$

which tends to a limit $d_1 = -(B+D) + \sqrt{B^2 + 2BD}$.

Consequently $e_{j,1} \rightarrow e_1 = D/d_1$.

Finally $f_{j,1} = (B - D f_{j-1,1})/d_{j,1}$. For large j this becomes

$$f_{j,1} = \frac{B - D f_{j-1,1}}{d_1}$$

which approaches $B/(d_1 + D)$ as j increases. Therefore $G_{j,1} \rightarrow G_1 = c_1 f_1$.

Now assume the truth of the lemma for $k < K$. For $k = K$ we have the following.

$$b_{j,K} \rightarrow B - G_{K-1} = b_K.$$

$$c_{j,K} \rightarrow D - G_{K-1} = c_K.$$

$$d_{j,K} \rightarrow -2(B+D) + 2G_{K-1} - b_K f_{K-1} + c_K^2/d_K.$$

This is in the same form as (5.1) with only the constants changed. Thus

$d_{j,K}$ tends to

$$d_K = -(B+D) + G_{K-1} - \frac{1}{2} b_K f_{K-1} + \frac{1}{2} \sqrt{(-2(B+D) + 2G_{K-1} - b_K f_{K-1})^2 - 4c_K^2}$$

$$\text{and } e_{j,K} \rightarrow e_K = \frac{D - b_K e_{K-1}}{d_K}.$$

Finally $f_{j,K} \rightarrow f_K = B/(d_K + c_K)$ and $G_{j,K} \rightarrow G_K = c_K f_K$ which completes the proof.

The next lemma will show that the quantities b_k and d_k of the previous lemma are increasing in magnitude as k increases if $0 > B > D$. Lemma 5.4 then shows that the b_k are increasing in magnitude for any $B, D < 0$. This result is used in Lemma 5.5 to prove that as j and k increase the quantities $b_{j,k}$ and $d_{j,k}$ are approaching limits b and d respectively.

Lemma 5.3: Suppose $0 > B > D$. Let b_k and d_k be as in Lemma 5.2. Then the b_k are decreasing and the d_k are increasing as k increases.

Proof: The proof will again be by an induction.

$$b_1 = 0.$$

$$c_1 = D.$$

$$d_1 = -(B+D) + \sqrt{B^2 + 2BD}.$$

$$f_1 = B/(d_1 + D) = (B + \sqrt{B^2 + 2BD})/2D.$$

and

$$G_1 = c_1 f_1 = (B + \sqrt{B^2 + 2BD})/2.$$

To show that $b_2 < b_1$ and $d_2 > d_1$ we have the following relations.

$$b_2 = B - G_1 = (B - \sqrt{B^2 + 2BD})/2 < 0 = b_1 ,$$

$$c_2 = D - G_1 = (2D - B - \sqrt{B^2 + 2bD})/2 ,$$

and

$$d_2 = -(B+D) + G_1 - \frac{1}{2} b_2 f_1 + \frac{1}{2} \{ [-2(B+D) + 2G_1 - b_2 f_1]^2 - 4c_2^2 \}^{1/2} .$$

Substituting the above expressions for f_1 , G_1 , b_2 and c_2 yields

$$d_2 = \frac{-B}{4} - D + \frac{1}{2} \sqrt{B^2 + 2BD} + \frac{1}{2} \left(-\frac{3}{4} B^2 + 6BD - 3B \sqrt{B^2 + 2BD} \right) \frac{1}{2} .$$

Since $B > D$,

$$-3B \sqrt{B^2 + 2BD} > -3B \sqrt{B^2 + 2BD} + \frac{12B^2}{4} + B^2 + 2BD - 6BD. \quad (5.2)$$

Collecting terms taking the square root and multiplying by $1/2$ we obtain

$$\frac{1}{2} \left(-\frac{3}{4} B^2 + 6BD - 3B \sqrt{B^2 + 2BD} \right)^{1/2} > -\frac{3}{4} B + \frac{1}{2} \sqrt{B^2 + 2BD} .$$

Using this inequality and the last expression for d_2 yields

$$d_2 > -B - D + \sqrt{B^2 + 2BD} = d_1 .$$

The following relations, which come from substituting the limiting values of $b_{j,k}$ and $d_{j,k}$ into (2.7) will be required in the general step of the induction.

$$b_k = B \left(1 + \frac{b_{k-1} - B + D}{d_{k-1}} \right)^{-1} \quad k = 2, 3, \dots \quad (5.3)$$

and

$$d_k = -2(D + b_k) - \frac{b_k^2}{d_{k-1}} - \frac{(b_k - B + D)^2}{d_k} \quad k = 2, 3, \dots \quad (5.4)$$

Note that $b_k < B < 0$ implies

$$0 < 1 + \frac{b_{k-1} - B + D}{d_{k-1}} < 1$$

and thus

$$-1 < \frac{b_{k-1} - B + D}{d_{k-1}} < 0. \quad (5.5)$$

Before starting the inductive step, it is necessary to show that $b_3 < b_2$. This inequality holds if

$$\left(1 + \frac{b_2 - B + D}{d_2}\right)^{-1} > \left(1 + \frac{b_1 - B + D}{d_1}\right)^{-1}$$

or equivalently if

$$\frac{b_2 - B + D}{d_2} < \frac{-B + D}{d_1} . \quad (5.6)$$

Replacing b_2 and d_1 by $(B - \sqrt{B^2 + 2BD})/2$ and $-(B + D) + \sqrt{B^2 + 2BD}$ (5.6) becomes

$$\frac{-3BD - 2D^2 + 3D}{2d_2} \frac{B^2 + 2BD}{B^2 + 2BD} < -B + D .$$

Substituting for d_2 and collecting terms we get

$$(B + 2D) \left(-\frac{B}{2} + \sqrt{B^2 + 2BD}\right) - \frac{7}{2} BD < (D - B) \left(-\frac{3}{4} B^2 + 6BD - 3B\sqrt{B^2 + 2BD}\right)^{1/2} . \quad (5.7)$$

This will be true if

$$(B^2 + 4BD + 4D^2) \left(\frac{B^2}{4} - B\sqrt{B^2 + 2BD} + B^2 + 2BD\right) - 7BD(B + 2D)\sqrt{B^2 + 2BD}$$

is greater than

$$(D^2 - 2BD + B^2) \left(-\frac{3}{4} B^2 + 6BD - 3B\sqrt{B^2 + 2BD}\right) .$$

(That $D - B$ is negative was used in obtaining this condition.) Multiplying out the factors of the above inequality leads to

$$2B^4 - 1/2 B^3 D + 25\frac{3}{4} B^2 D^2 + 2BD^3 - B\sqrt{B^2 + 2BD} (-2B^2 + 15D^2 + 17BD) > 0$$

which is obviously true. Thus (5.7) is valid and therefore (5.6) is valid and $b_3 < b_2$.

We now start the inductive step. Assume $b_k < b_{k-1}$ and

$d_{k-1} > d_{k-2}$. Let $\delta_k = b_{k-1} - b_k > 0$, and let $\epsilon_k = d_k - d_{k-1}$. From (5.4)

$$d_k = -2(D+b_k) - \frac{b_k^2}{d_{k-1}} - \frac{(b_k - B + D)^2}{d_k}.$$

Replacing b_k with $b_{k-1} - \delta_k$ and d_k with $\epsilon_k + d_{k-1}$ introduces $-2(D+b_{k-1})$.

Then using (5.4) and the definition of ϵ_k yields

$$\epsilon_k = 2\delta_k + \frac{b_{k-1}^2}{d_{k-2}} - \frac{(b_{k-1} - \delta_k)^2}{d_{k-1}} + \frac{(b_{k-1} - B + D)^2}{d_{k-1}} - \frac{(b_{k-1} - \delta_k - B + D)^2}{d_{k-1} + \epsilon_k} \quad (5.8)$$

Since $d_{k-2} < d_{k-1}$ by assumption,

$$\begin{aligned} \epsilon_k &> 2\delta_k + \frac{b_{k-1}^2 + 2\delta_k b_{k-1} - \delta_k^2 + B^2 + D^2 - 2Bb_{k-1} + 2Db_{k-1} - 2BD}{d_{k-1}} - \frac{(b_{k-1} - \delta_k - B + D)^2}{d_{k-1} + \epsilon_k} \\ &= \frac{2\delta_k}{d_{k-1}} (d_{k-1} + b_k - B + D + b_{k-1}) + \frac{\epsilon_k (b_k - B + D)^2}{d_{k-1} d_k}. \end{aligned} \quad (5.9)$$

Now from part (i) of Lemma 5.1 we have $d_{k-1}^{B+D-b_{k-1}} e_{k-2}^{-c_{k-1}} f_{k-1} > 0$.

But

$$-b_{k-1} e_{k-1} = -c_{k-2} f_{k-2} = b_{k-1}^{-B},$$

and

$$-c_{k-1} f_{k-1} = b_k^{-B}.$$

Hence

$$d_{k-1}^{B+D+b_{k-1}-B+b_k-B} > 0$$

and

$$d_{k-1}^{D-B+b_k+b_{k-1}} > 0. \quad (5.10)$$

Now (5.9) and (5.10) together with $\delta_k > 0$ and $d_{k-1} > 0$ yield

$$\epsilon_k > \frac{\epsilon_k (b_k - B + D)^2}{d_{k-1} d_k}.$$

Furthermore (5.10) implies $|(b_k - B + D)/d_{k-1}| < 1$ and (5.5) implies

$$|(b_k - B + D)/d_k| < 1. \quad \text{Hence } e_k > 0 \text{ and therefore } d_k > d_{k-1}.$$

It remains necessary to show that $b_{k+1} < b_k$. We will show that $e_k < e_{k-1}$, where e_k is defined as in Lemma 5.2 (This in turn requires $f_{k-1} < f_{k-2}$, (5.12) below).

From (5.3)

$$b_{k+1} = B \left[\frac{1}{\frac{b_k - B + D}{1 + \frac{1}{d_k}}} \right].$$

But

$$\frac{b_k - B + D}{d_k} = \frac{c_k}{d_k} = e_k.$$

$$\text{Thus } b_{k+1} = B \left(\frac{1}{1 + e_k} \right). \quad (5.11)$$

Therefore if $e_k < e_{k-1}$ then

$$b_{k+1} < B/(1 + e_{k-1}) = b_k.$$

Since $b_k < b_{k-1}$, (5.11) implies $e_{k-1} < e_{k-2}$. Next from the definition of $d_{j,k}$ along with

$$-2D + c_{k-1}f_{k-1} + b_k f_{k-1} = -2c_k = -2d_k e_k$$

we arrive at

$$d_k + d_k e_k^2 = -2B - 2d_k e_k - b_k f_{k-1}.$$

Thus

$$(1 + e_k)^2 = (-2B - b_k f_{k-1})/d_k.$$

But $d_k > d_{k-1}$ and $b_k < b_{k-1}$. If in addition

$$f_{k-1} < f_{k-2} \quad (5.12)$$

Then $-2B-b_k f_{k-1} < -2B-b_{k-1} f_{k-2}$ and

$$(1+e_k)^2 < \frac{-2B-b_{k-1} f_{k-2}}{d_{k-1}} = (1+e_{k-1})^2 . \quad (5.13)$$

But $1+e_k$ and $1+e_{k-1}$ are positive so that (5.13) implies

$$1+e_k < 1+e_{k-1}$$

or

$$e_k < e_{k-1} . \quad (5.14)$$

To see that $f_{k-1} < f_{k-2}$ we write

$$c_k + c_{k-1} f_{k-1} = D = c_{k-1} + c_{k-2} f_{k-2} .$$

Since $c_k - c_{k-1} = b_k - b_{k-1}$, the last equation implies

$$c_{k-1} - \delta_k + c_{k-1} f_{k-1} = c_{k-1} + (c_{k-1} + \delta_{k-1}) f_{k-2} .$$

$$c_{k-1} (f_{k-1} - f_{k-2}) = \delta_k + \delta_{k-1} f_{k-2} .$$

$$b_{k-1} (f_{k-1} - f_{k-2}) + (-B+D) (f_{k-1} - f_{k-2}) = \delta_k + \delta_{k-1} f_{k-2} .$$

$$b_{k-1} (f_{k-1} - f_{k-2}) = f_{k-2} (D-B+\delta_{k-1}) + (-D+B) f_{k-1} + \delta_k .$$

$$f_{k-2} (-b_{k-1} + B-D-\delta_{k-1}) = f_{k-1} (-D+B-b_{k-1}) + \delta_k .$$

$$f_{k-2} (-b_{k-2} + B-D) > f_{k-1} (-b_{k-1} + B-D) .$$

But $c_k = b_k - B + D$ so

$$f_{k-2} > f_{k-1} \frac{-c_{k-1}}{-c_{k-2}} > f_{k-1} .$$

Hence (5.12) holds and implies (5.14) which implies $b_{k+1} < b_k$. This

completes the proof.

If we remove the condition $B > D$ we can still prove $b_{k+1} < b_k$

for all k although it is no longer true that $d_{k+1} > d_k$. The restriction $B > D$ was used only in (5.2) to show $d_2 > d_1$ and in (5.7) to show $b_3 < b_2$. If $D \geq B$ then (5.7) is trivially true and thus $b_3 < b_2$. Furthermore if for some k , $d_k \leq d_{k-1}$ and $b_k < b_{k-1}$, then from (5.3)

$$b_{k+1} = B \left(1 + \frac{b_k - B + D}{d_k} \right)^{-1} < B \left(1 + \frac{b_{k-1} - B + D}{d_{k-1}} \right)^{-1} = b_k.$$

That is, $b_{k+1} < b_k$ whether $d_k \leq d_{k-1}$ or $d_k > d_{k-1}$. Thus the previous lemma handled the more difficult case of the following lemma.

Lemma 5.4: The b_k as defined in Lemma 5.2 are decreasing.

Lemma 5.4 can now be used to prove all the elements of the matrix B approach limits.

Lemma 5.5: Let $B_{j,k} \equiv B$ and $D_{j,k} \equiv D$. The quantities $b_{j,k}$, $c_{j,k}$, $d_{j,k}$, $e_{j,k}$ and $f_{j,k}$ as defined in (2.7) approach limiting values as j and k increase. We will call these values b, c, d, e , and f respectively.

Proof: Since all elements can be defined in terms of $b_{j,k}$ and $d_{j,k}$, we will prove only that $b_{j,k} \rightarrow b$ and $d_{j,k} \rightarrow d$.

By Lemma 5.4, the b_k are increasing in magnitude as k increases.

It is sufficient to show that the b_k are bounded below for then they must approach a limit, b . Moreover the d_k must approach a limit since

$$b_k = B \left(1 + \frac{b_{k-1} - B + D}{d_{k-1}} \right)^{-1}.$$

Let $d_{\min} = \min \{d_k\}$ and let $\bar{b}_1 = 0$, $\bar{b}_2 = b_2$ and

$$\bar{b}_k = B \left(\frac{d_{\min}}{d_{\min} + \bar{b}_{k-1} - B + D} \right).$$

Now suppose $\bar{b}_k \leq b_k$ then

$$\begin{aligned} b_{k+1} &= B \frac{d_{\min}}{d_{\min} + \bar{b}_k - B + D} \\ &\leq B \frac{d_{\min}}{d_{\min} + b_k - B + D} \end{aligned}$$

since the denominator is positive and increased. The fraction is greater than one so that adding the same positive number to the numerator and denominator decreases it. Hence

$$\bar{b}_{k+1} \leq B \frac{d_k}{d_k + b_k - B + D} = b_{k+1}.$$

Thus $\bar{b}_k \leq b_k$ for all k . Furthermore the $\bar{b}_k$ decrease to

$$\lim_{k \rightarrow \infty} \bar{b}_k = \frac{-d_{\min} + B - D + \sqrt{(d_{\min} - B + D)^2 + 4Bd_{\min}}}{2}.$$

The b_k , therefore, are bounded below and the proof is complete.

The next theorem establishes values for the limits of b_k and d_k .

Theorem 5.1: If b_k and d_k are as in Lemma 5.2 then

$$\lim_{k \rightarrow \infty} b_k = B - \sqrt{BD} \quad \text{and} \quad \lim_{k \rightarrow \infty} d_k = -B - D + 2\sqrt{BD}.$$

Proof: From Lemma 5.5 we have that b_k and d_k approach limits b and d respectively. In addition these limits satisfy (5.3) and (5.4) since the b_k and d_k satisfy them. Hence

$$b = \frac{Bd}{d+b-B+D} = B \left(1 + \frac{b-B+D}{d} \right)^{-1} \quad (5.15)$$

and

$$d = -2(D+b) - \frac{1}{d} (b^2 + (b-B+D)^2).$$

Therefore

$$d^2 + 2d(D+b) + b^2 + (b-B+D)^2 = 0. \quad (5.16)$$

From (5.15)

$$b^2 = B^2 \left(1 + \frac{2(b-B+D)}{d} + \frac{(b-B+D)^2}{d^2} \right)^{-1}.$$

Using (5.16) to replace $(b-B+D)^2/d^2$ yields

$$\begin{aligned} b^2 &= B^2 \left(1 + \frac{2(b-B+D)}{d} - 1 - \frac{2(D+b)}{d} - \frac{b^2}{d^2} \right)^{-1} \\ &= \frac{B^2 d^2}{-2Bd - b^2}. \end{aligned}$$

Consequently

$$-b^4 - b^2(2Bd) - B^2 d^2 = 0.$$

and

$$b^2 = \frac{2Bd \pm \sqrt{4B^2 d^2 - 4B^2 d^2}}{-2} = -Bd.$$

Since $b < 0$, the last equality implies

$$b = -\sqrt{-Bd}. \quad (5.17)$$

Substituting (5.17) into (5.16) yields

$$d^2 + 2d(D - \sqrt{-Bd}) - 2Bd - 2\sqrt{-Bd}(-B+D) + (-B+D)^2 = 0.$$

Collecting terms we get

$$(d - B + D)^2 - 2\sqrt{-Bd}(d - B + D) = 0.$$

Hence

$$-2b = 2\sqrt{-Bd} = d - B + D. \quad (5.18)$$

Squaring both sides of the second equality and again collecting terms

$$0 = d^2 + B^2 + D^2 + 2Bd + 2dD - 2BD.$$

Therefore

$$\begin{aligned} d &= \frac{-2B - 2D \pm \sqrt{4B^2 + 4D^2 + 8BD + 8BD - 4B^2 - 4D^2}}{2} \\ &= -B - D \pm 2\sqrt{BD}. \end{aligned}$$

Note that if $d = -B - D - 2\sqrt{BD}$ then (5.18) implies $b = B + \sqrt{BD}$. But

$$b \leq b_k < B.$$

Therefore we must have

$$d = -B - D + 2\sqrt{BD}$$

and

$$b = B - \sqrt{BD}.$$

This completes the proof.

Corollary 5.1.1: If $B_{j,k} \equiv B$ and $D_{j,k} \equiv D$ then $G_{j,k}$ of (2.7) tends to $G = \sqrt{BD}$ as j and k increase.

Proof: This is obvious from the theorem.

5.2 An Estimate Of $\lambda_{\min}$

In this section an estimate of $\lambda_{\min}$ to be used in the algorithm of Chapter 4 is developed. We begin by examining the case

$$B_{j,k} \equiv 1 \text{ and } D_{j,k} \equiv 1 \text{ except } B_{j,1} = D_{1,k} = 0.$$

Then some more general cases are considered.

It will be shown that for $B = D = 1$, $\lambda_{\min} ((A+B)^{-1}A)$ approaches $1/2$ from above as the meshsize, h , of the grid used in deriving (2.2) from (2.1), tends to zero. We begin with the following lemma which will be used to show $\lambda_{\min} ((A+B)^{-1}A) \geq 1/2$ for all $h > 0$.

Lemma 5.6: Let A be as in (2.4) with $B_{j,k} \equiv D_{j,k} \equiv 1$ except $B_{j,1} = D_{1,k} = 0$ $j, k = 1, 2, \dots, n$. Let $B = (\hat{b}_{i,j})$ be the auxiliary matrix derived from (2.7) using $\alpha = 1$. Then for all $x \neq 0$, $\langle Ax, x \rangle > \langle Bx, x \rangle$.

Proof: Bracha has shown in [1] that if $A = (a_{i,j})$ is an N by N matrix in the form of (2.4) and $B = (\hat{b}_{i,j})$ is given by (2.7) then

$$\begin{aligned} \langle Ax, x \rangle &= \sum_{i=1}^{N-1} -a_{i,i+1} |x_{i+1} - x_i|^2 + \sum_{i=1}^{N-n} -a_{i,i+n} |x_{i+n} - x_i|^2 \\ &+ \sum_{i=1}^N (a_{i,i} + a_{i,i+1} + a_{i,i+n} + a_{i-1,i} + a_{i-n,i}) |x_i|^2 \end{aligned}$$

and

$$\begin{aligned} \langle Bx, x \rangle &= - \sum_{i=2}^{N-1} \hat{b}_{i,i+1} |x_{i+n} - x_i|^2 + \sum_{i=2}^{N-n} \hat{b}_{i,i+n} |x_i|^2 \\ &- \sum_{i=n+1}^{N-1} \hat{b}_{i,i+1} |x_{i+1} - x_i|^2 + \sum_{i=n+1}^{N-1} \hat{b}_{i,i+1} |x_i|^2 \end{aligned} \quad (5.19)$$

$$- \sum_{i=1}^{N-n-1} \hat{b}_{i+1,i+n} |x_{i+1} - x_{i+n}|^2 + \sum_{i=1}^{N-n-1} \hat{b}_{i+1,i+n} (|x_{i+1}|^2 + |x_{i+n}|^2),$$

where $N = n^2$.

The second and fourth sums may be rewritten as

$$\sum_{i=1}^{N-n-1} \hat{b}_{i+1,i+n+1} |x_{i+1}|^2 \quad (5.20)$$

and

$$\sum_{i=1}^{N-n-1} \hat{b}_{i+n,i+n+1} |x_{i+n}|^2 \quad (5.21)$$

Now from the definition of B

$$\hat{b}_{i+1,i+n+1} = -\alpha \hat{b}_{i+1,i+n}$$

and

$$\hat{b}_{i+n,i+n+1} = -\alpha \hat{b}_{i+n,i+1} = -\alpha \hat{b}_{i+1,i+n}.$$

Hence for $\alpha = 1$, (5.20) and (5.21) become

$$- \sum_{i=1}^{N-n-1} \hat{b}_{i+1,i+n} |x_{i+1}|^2 \text{ and } - \sum_{i=1}^{N-n-1} \hat{b}_{i+1,i+n} |x_{i+n}|^2$$

respectively. Therefore,

$$\begin{aligned} \langle (A-B)x, x \rangle &= \sum_{i=1}^{N-1} -a_{i,i+1} |x_{i+1} - x_i|^2 + \sum_{i=n+1}^{N-1} \hat{b}_{i,i+1} |x_{i+1} - x_i|^2 \\ &\quad + \sum_{i=1}^{N-n} -a_{i,i+n} |x_{i+n} - x_i|^2 + \sum_{i=2}^{N-n} \hat{b}_{i,i+n} |x_{i+n} - x_i|^2 \end{aligned} \quad (5.23)$$

$$\begin{aligned}
& + \sum_{i=1}^N (a_{i,1} + a_{i,i+1} + a_{i,i+n} + a_{i-n,i} + a_{i-1,i}) |x_i|^2 \\
& + \sum_{i=1}^{N-n-1} \hat{b}_{i+1,i+n} |x_{i+1} - x_{i+n}|^2 . \\
& = \sum_{i=1}^{N-1} (-a_{i,i+1} + \hat{b}_{i,i+1}) |x_{i+1} - x_i|^2 \\
& + \sum_{i=1}^{N-n} (-a_{i,i+n} + \hat{b}_{i,i+n}) |x_{i+n} - x_i|^2 \\
& + \sum_{i=1}^N (a_{i,i} + a_{i,i+1} + a_{i,i+n} + a_{i-1,i} + a_{i-n,i}) |x_i|^2 \\
& + \sum_{i=1}^{N-n-1} \hat{b}_{i+1,i+1} |x_{i+1} - x_{i+1}|^2 .
\end{aligned} \tag{5.23}$$

From Lemma 5.5 and Theorem 5.1 it is seen that $|b_{i,j}| < 1$ and from

Lemma 5.1 it is seen that $b_{i+1,i+n} > 0$. Hence all the coefficients

in (5.23) are positive since $a_{i,j} = -1$ for $i \neq j$. Therefore $\langle (A-B)x, x \rangle > 0$

proving the lemma.

Theorem 5.2: If h , A and B are as in Lemma 5.6, $\lambda_{\min}((A+B)^{-1}A) > 1/2$

and $\lambda_{\min}((A+B)^{-1}A) \rightarrow 1/2$ as $h \rightarrow 0$.

Proof: From Lemma 5.6 we have

$$\langle (A+B)x, x \rangle < 2 \langle Ax, x \rangle .$$

Thus $\frac{\langle Ax, x \rangle}{\langle (A+B)x, x \rangle} > \frac{\langle Ax, x \rangle}{2 \langle Ax, x \rangle} = 1/2$. Using (3.6) yields

$$\lambda_{\min}((A+B)^{-1}A) > 1/2 .$$

To establish the remainder of the theorem consider the vector

$$x = (x_{1,1}, \dots, x_{n,n})^T \text{ such that } x_{j,k} = \begin{cases} 1 & \text{if } j+k=n, j > J \text{ and } k > K \\ 0 & \text{otherwise} \end{cases} \quad (5.24)$$

where J and K are chosen so large that the elements of B have nearly constant value. That is, if row i of B corresponds to the point (j,k)

with $j > J$ and $k > K$ then $b_{i,i-n} \cong -1$; $b_{i,i-n+1} \cong 1$; $b_{i,i-1} \cong -1$;

$b_{i,i} \cong 2$; $b_{i,i+1} \cong -1$; $b_{i,i+n-1} \cong 1$ and $b_{i,i+n} \cong -1$.

$$\langle Ax, x \rangle = (n-K-1)4 \text{ and } \langle Bx, x \rangle \cong (n-K-3)4+6.$$

$$\frac{\langle Ax, x \rangle}{\langle (A+B)x, x \rangle} \cong \frac{(n-K-1)4}{2(n-K-1)4-2} = \frac{1}{2 - \frac{2}{(n-K-1)}}. \quad (5.25)$$

as h goes to 0, (i.e., as n goes to infinity) the quantity in (5.25) tends to $1/2$ from the above. Using (3.6) again completes the proof.

The argument used to prove the second half of the above theorem can be used to prove a similar result for a more general matrix A . Suppose A is derived from a differential operator in (2.1) with constant $a_1(x)$ and $a_2(x)$. Then A will have constant diagonals $a_{i,i-n} \equiv B$, $a_{i,i-1} \equiv D$ and $a_{i,i} \equiv -2(B+D)$ except for some zero elements on rows corresponding to boundary points. Furthermore the elements of the auxiliary matrix B tend to the following values: $b_{i,i-n} \rightarrow -\sqrt{BD}$; $b_{i,i-(n-1)} \rightarrow \sqrt{BD}$; $b_{i,i-1} \rightarrow -\sqrt{BD}$ and $b_{i,i} \rightarrow 2\sqrt{BD}$. Thus if X is defined as in (5.24) then

$$\begin{aligned} \frac{\langle Ax, x \rangle}{\langle (A+B)x, x \rangle} &\cong \frac{(n-K-1) (-2) (B+D)}{(n-K-1) (-2) (B+D) + (n-K-3)4(\sqrt{BD}) + 6\sqrt{BD}} \\ &= \frac{-2(n-K-1) (B+D)}{2(n-K-1) (-B-D + \sqrt{BD}) - 2\sqrt{BD}} \end{aligned}$$

$$\cong \frac{-(B+D)}{-(B+D) + 2\sqrt{BD}} = \frac{1}{2 + \frac{2\sqrt{BD}}{-(B+D)}}.$$

(3.6) and the above yield the following theorem.

Theorem 5.3: If A is in the form of (2.6) with $B_{j,k} \equiv B$ and $D_{j,k} \equiv D$ except $B_{j,1} = 0$ and $D_{1,k} = 0$ and if B is the auxiliary matrix defined in (2.10) using $\alpha = 1$, then

$$\lim_{h \rightarrow 0} \lambda_{\min}((A+B)^{-1}A) \leq \frac{1}{1 + \frac{2\sqrt{BD}}{-(B+D)}}, \quad (5.26)$$

where h is the meshsize of the grid used in the discretization of the differential problem.

There is no lemma corresponding to Lemma 5.6 which implies the inequality in (5.26) can be reversed. That is, the quantity in (5.26) is not $\lim_{h \rightarrow 0} \lambda_{\min}((A+B)^{-1}A)$. However this value has proved to be effective as the initial left endpoint, a_1 , in the algorithm of Chapter 4. This value is suggested.

The result of the previous theorem can be generalized to a matrix A which is block tridiagonal with constant diagonals within each block. The result is of little value however since again there is no lemma which like Lemma 5.6 establishes a lower bound for the eigenvalues. Furthermore, the value, which arises from this modification of the argument leading to Theorem 5.3, is so crude an estimate of $\lambda_{\min}$ that it is not worth calculating for use in the algorithm.

As mentioned in Chapter 4 the initial left endpoint, a_1 , to use in the algorithm should be calculated according to Theorem 5.3 when possible. If the Theorem is not applicable then $a_1 \in [1/4, 1/2]$ is sufficient.

LIST OF REFERENCES

- [1] Bracha, A., "An Implicit Method for the Solution of Elliptic Partial Differential Equations", Department of Computer Science, University of Illinois, Urbana, Illinois, 61801, Report No. 392, April, 1970.
- [2] Bracha, A. "A Symmetric Procedure for the Solution of Elliptic Boundary Value Problems", Department of Computer Science, University of Illinois, Urbana, Illinois, Report No. 440, April, 1971.
- [3] Carré, B. A., "The Determination of the Optimum Accelerating Factor for Successive Over-Relaxation", The Computer Journal, Vol. 4, April, 1961.
- [4] Dupont, T., "A Factorization Procedure for the Solution of Elliptic Difference Equations", SIAM Journal On Numerical Analysis, December, 1968.
- [5] Dupont, T., R. P. Kendall, and H. H. Rachford, Jr., "An Approximate Factorization Procedure for Solving Self-Adjoint Elliptic Difference Equations", SIAM Journal on Numerical Analysis, September, 1968.
- [6] D'Yakonov, E. G., "On the Solution of Some Elliptic Difference Equations", Moscow State University, U.S.S.R., 1969.
- [7] Engeli, M., "Over-Relaxation and Related Methods"(Chapter IV of Refined Iterative Methods for Computation of the Solution and the Eigenvalues of Self-Adjoint Boundary Value Problems, by M. Engeli, Mitteilungen aus dem Institut für angewandte Mathematik an der Eidgenössischen Technischen Hochschule Zürich, Nr. 8). Birkhäuser Verlag, Basel/Stuttgart, 1959.
- [8] Frank, W. L., "Solution of Lemar Systems by Richardson's Method", J. Assoc. Compu. March, Vol. 7, July, 1960.
- [9] Frankel, S. P., "Convergence Rates of Iterative Treatments of Partial Differential Equations", MTAC, Vol. 4, 1950.
- [10] Hageman, L. A. and R. B. Kellogg, "Estimating Optimum Overrelaxation Parameters", Math. Comp., Vol. 22, 1968.
- [11] Kim, Y. J., Private Communication, 1971.
- [12] Rigler, A. K., "Estimation of the Successive Over-Relaxation Factor", Math. Comp., Vol. 19, 1965.
- [13] Saul'yev, V. K., "Integration of Equations of Parabolic Type by the Method of Nets", Pergamon Press, 1964.

- [14] Sheldon, J. W., "On the Numerical Solutions of Elliptic Difference Equations", MTAC, Vol. 9, 1955.
- [15] Stone, H. L., Private Communication, April, 1969.
- [16] Stone, H. L., "Iterative Solution of Implicit Approximations of Multidimensional Partial Differential Equations", SIAM Journal On Numerical Analysis, September, 1968.
- [17] Taranto, R., "Numerical Studies of Stones' Symmetric Factorization and the Iteration Parameters α and τ ", Department of Computer Science, University of Illinois, Urbana, Illinois, 61801, Report No. 423, January, 1971.
- [18] Varga, R. S., "A Comparison of the Successive Overrelaxation Method and Semi-Iterative Methods Using Thebychev Polynomials", T. Soc. Indust. Appl. Math., Vol. 5, June, 1957.
- [19] Varga, R. S., Matrix Iterative Analysis, Prentice-Hall, Englewood Cliffs, New Jersey, 1962.
- [20] Wackspress, E. L., Iterative Solution of Elliptic Systems, Prentice-Hall, Englewood Cliffs, New Jersey, 1966.
- [21] Wilkenson, J. H., The Algebraic Eigenvalue Problem, Clarendon Press, Oxford, 1965.
- [22] Wilkenson, J. H., "The Calculation of the Latent Roots and Vectors of Matrix on the Pilot Model of the ACE", Proc. Camb. Phil. Soc. 50, 1954.
- [23] Wrigley, H. E. "Accelecatig the Jacobi Method for Solving Simultaneous Equations by Thebyshev Extrapolation when then Eigenvalues of the Iteration Matrix are Complex", The Computer Journal, Vol, 6, July, 1963.

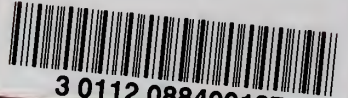
VITA

Martin Allen Diamond was born in Los Angeles, California in September 1945. After receiving a B.A. in Mathematics from Arizona State University in 1967, he was granted an NDEA fellowship and continued studying at that university until August 1968 when he received an M.A. in Mathematics. Since September 1968 he has been a teaching assistant and research assistant in the Department of Computer Science of the University of Illinois at Urbana-Champaign.

JUN 7 1972



UNIVERSITY OF ILLINOIS-URBANA
510.84 IL6R no. C002 no. 487-492(1971)
Digital computer internal report /



3 0112 088400137